Anaphora With Non-nominal Antecedents in Computational Linguistics: a Survey

Varada Kolhatkar
Simon Fraser University
Department of Linguistics
varada.kolhatkar@gmail.com

Adam Roussel
Ruhr-Universität Bochum
Sprachwissenschaftliches Institut
roussel@linguistics.rub.de

Stefanie Dipper
Ruhr-Universität Bochum
Sprachwissenschaftliches Institut
dipper@linguistics.rub.de

Heike Zinsmeister
Universität Hamburg
Institut für Germanistik
heike.zinsmeister@uni-hamburg.de

*This article provides an extensive overview of the literature related to the phenomenon of non-nominal-antecedent anaphora (also known as abstract anaphora or discourse deixis), a type of anaphora in which an anaphor like "that" refers to an antecedent (marked in boldface) that is syntactically non-nominal, such as the first sentence in "**It's way too hot here**. **That**'s why I'm moving to Alaska." Annotating and automatically resolving these cases of anaphora is interesting in its own right because of the complexities involved in identifying non-nominal antecedents, which typically represent abstract objects such as events, facts, and propositions. There is also practical value in the resolution of non-nominal-antecedent anaphora, as this would help computational systems in machine translation, summarization, and question answering, as well as, conceivably, any other task dependent on some measure of text understanding.*

Most of the existing approaches to anaphora annotation and resolution focus on nominal-antecedent anaphora, classifying many of the cases where the antecedents are syntactically

Submission received: 5 March 2018; revised version received: 19 June 2018; accepted for publication: 22 June 2018.

doi:10.1162/COLI_a_00327

non-nominal as non-anaphoric. There has been some work done on this topic, but it remains scattered and difficult to collect and assess. With this article, we hope to bring together and synthesize work done in disparate contexts up to now in order to identify fundamental problems and draw conclusions from an overarching perspective. Having a good picture of the current state of the art in this field can help researchers direct their efforts to where they are most necessary.

Because of the great variety of theoretical approaches that have been brought to bear on the problem, there is an equally diverse array of terminologies that are used to describe it, so we will provide an overview and discussion of these terminologies. We also describe the linguistic properties of non-nominal-antecedent anaphora, examine previous annotation efforts that have addressed this topic, and present the computational approaches that aim at resolving non-nominal-antecedent anaphora automatically. We close with a review of the remaining open questions in this area and some of our recommendations for future research.

1. Introduction

This article addresses a particular subset of the phenomenon of **anaphora**, which is central to natural language understanding. Anaphora refers to a relation between two linguistic entities, an **anaphor** and an **antecedent**, in which the interpretation of the anaphor is contingent upon the meaning of the antecedent (Huddleston and Pullum 2002, page 1,453).[1] Example (1) shows a typical case, in which *she* is the anaphor and *Maya* is the antecedent.[2]

(1) **Maya** went to the farmer's market this morning. **She** bought fresh berries to make delicious smoothies.

There are several ways one might categorize instances of anaphora. A simple means of categorization, for instance, would involve examining the syntactic form of either the anaphor or its antecedent, contrasting pronouns and full noun phrases as anaphors, or nominal and non-nominal antecedents, respectively (by **nominal**, we mean expressions in the form of full noun phrases [NPs] or pronouns). The cases of anaphora in which the antecedent is of a non-nominal syntactic type, as in Example (2), present special challenges for computational approaches because of the variety of potential antecedents of this type, which are not always clearly delimited and identifiable stretches of text. Such cases of anaphora with non-nominal antecedents are the topic of this article.

(2) The municipal council had to decide **whether to balance the budget by raising revenue or cutting spending**. The council had to come to a resolution by the end of the month. **This issue** was dividing communities across the country.

1 Note that the terms *anaphor* and *antecedent* refer to textual entities whereas *anaphora* refers to a relation.

2 In all examples in this article, the anaphoric expressions are shown in boldface and are underlined, and their antecedents (i.e., the linguistic constituents most closely representing the intended interpretation of the anaphor) are shown in boldface, if not stated otherwise. Examples without a specified source are constructed examples.

Previous work in the field is spread across a plethora of terminologies, such as *abstract anaphora* (Asher 1993; Navarretta 2007; Dipper et al. 2011), *discourse deixis* (Webber 1988; Eckert and Strube 2000; Byron 2004; Recasens 2008), *indirect anaphora* (Gundel, Hedberg, and Zacharski 2004; Botley 2006), and *complex anaphora* (Consten, Knees, and Schwarz-Friesel 2007), to name a few. The majority of these terminologies highlight the semantic aspects of the phenomenon—for example, whether the antecedent refers to an abstract object or not. Though each of these terminologies reflects slight variations in how the authors define these phenomena, they all have in common that they describe anaphora where antecedents are not the usual noun phrases (e.g., as in Example (2)) and where referents are abstract, proposition-like entities. We focus on the syntactic aspect of the phenomenon and adopt the name **non-N**ominal-**A**ntecedent (**non-NA**) anaphora for two reasons. First, we consider the non-nominal syntactic type as a characteristic property of these proposition-like entities. And second, the non-nominal syntactic form of the antecedents is what presents a distinctive and challenging problem for computational approaches, because (a) the search space of non-NA candidates is large, (b) non-NAs are not always clearly delimited and identifiable stretches of discourse, and (c) the features that are typically used to resolve pronominal anaphora (gender, number, person, etc.) are not available for non-NAs. To contrast non-NA anaphora and make the scope of this article clearer, we also define the term **N**ominal-**A**ntecedent (**NA**) anaphora, which encompasses all cases where the antecedents are noun phrases (e.g., as in Example (1)).

This article is motivated in part by the importance of anaphora for many natural language processing (NLP) tasks, such as machine translation (Le Nagard and Koehn 2010; Hardmeier, Nakov, Stymne, Tiedemann, Versley, and Cettolo 2015), summarization (Steinberger, Kabadjov, Poesio, and Sanchez-Graillet 2005; Orăsan 2007), and question answering (Ahn, Jijkoun, Mishne, Müller, de Rijke, and Schlobach 2004; Quarteroni 2007; Vicedo and Ferrández 2008). Anaphoric relations are ubiquitous in all kinds of texts, and non-NA anaphora is especially prevalent in spoken dialogue. For instance, Eckert and Strube (2000) indicate that 22.6% of the anaphors in their corpus of spoken dialogue refer to non-NAs. Being able to accurately identify and resolve such instances could conceivably improve the performance of downstream NLP tasks. For instance, consider the following dialogue from Gundel, Hegarty, and Borthen (2003):

(3) A: I just ate three pieces of cake.
B: Can you repeat **that**.

Here, a dialogue system aware of non-NAs could generate a clarification request: What should be repeated, the *act* of eating three pieces of cake or the *statement*? Automatically generated summaries could likewise benefit, allowing errant instances of *this* or *that* in the summary, which would otherwise be next to incomprehensible for the reader, to be replaced with their more informative antecedents.

Surveys have been written for NA anaphora and the closely related phenomenon of coreference (e.g., Hirst 1981; Mitkov 2002; Ng 2010; Poesio, Ponzetto, and Versley 2010; Poesio, Stuckardt, and Versley 2016) but these surveys do not account for the more complicated anaphoric relations involving clausal and verbal antecedents. There is as of yet no survey of work on non-NA anaphora as such in computational linguistics, and because of the variety of theoretical approaches and terminology brought to bear on the problem, the existing work remains scattered and difficult to collect and assess. This article thus aims to fill this gap, bringing together and synthesizing work done in disparate contexts up to now. We will identify the fundamental questions associated

with such anaphoric relations, and present a survey of the major approaches that tackle them.[3]

First, we define the phenomenon of non-NA anaphora from syntactic and semantic perspectives (Section 2). As mentioned before, the phenomenon of non-NA anaphora has been discussed in the literature in different contexts with a number of terminologies, such as *abstract anaphora*, *discourse deixis*, *indirect anaphora*, and *complex anaphora*. There are slight variations in meaning between these terminologies because they highlight different aspects of the phenomenon. In Section 2, we will discuss the commonalities and differences between these terminologies.

Second, we will discuss the linguistic properties of non-NA anaphora (Section 3). We start with the expressions that signal non-NA anaphora in English, such as demonstrative pronouns, the personal pronoun *it*, and noun phrases such as *this issue* and *this fact*, also known as *shell noun phrases* (Schmid 2000). We then discuss the linguistic properties of non-NA anaphora as presented in the literature: lexical and semantic properties (e.g., Asher 1993; Hegarty, Gundel, and Borthen 2001), syntactic properties (Passonneau 1989; Asher 1993), discourse-level properties (e.g., Gundel, Hegarty, and Borthen 2003; Poesio and Modjeska 2005; Hedberg, Gundel, and Zacharski 2007), and constraints on the distance between anaphor and antecedent, either in terms of tokens or sentences (e.g., Schmid 2000), the number of edges between nodes in some discourse representation structure (e.g., Webber 1991; Asher 1993, 2008), or spatio-temporal proximity (e.g., Halliday and Hasan 1976).

Third, we discuss relevant annotation efforts and describe major resources in detail with respect to different aspects, including the goals of the annotation, the anaphoric expressions considered, and information on inter-annotator agreement (Section 4). Analysis of non-NA anaphora has been carried out in a variety of domains and registers, from news texts (Botley 2006; Kolhatkar, Zinsmeister, and Hirst 2013a; Uryupina et al. 2018), to academic articles (Kolhatkar and Hirst 2012), to narratives (Botley 2006; Uryupina et al. 2018), to spoken monologues (Botley 2006; Guillou et al. 2014), to spoken dialogues (Schiffman 1985; Eckert and Strube 2000; Byron 2003; Uryupina et al. 2018) or multiparty discussions (Müller 2008). We see three kinds of approaches to annotating non-NAs: annotating semantic types of referents (e.g., Gundel, Hedberg, and Zacharski 2002; Byron 2003), annotating linguistic antecedents (e.g., Eckert and Strube 2000; Artstein and Poesio 2006; Kolhatkar and Hirst 2012; Kolhatkar, Zinsmeister, and Hirst 2013a; Guillou et al. 2014), and annotating a representative verbal head that acts as a proxy for a clausal linguistic antecedent (e.g, Müller 2008).

Fourth, we discuss computational approaches to non-NA anaphora (Section 5). These approaches can be broadly categorized into two classes: The first includes rule-based approaches, which treat the resolution of non-NA anaphora as the problem of selecting the appropriate discourse entity from a discourse model based on the current discourse state and the predication context of the anaphor (e.g., Eckert and Strube 2000; Byron 2004; Pappuswamy, Jordan, and VanLehn 2005). The second class includes machine-learning approaches, which focus on the problem of automatically identifying the linguistic

3 Note that this survey will focus on non-NA anaphora in English, because the majority of existing work also focuses on English. Work from theoretical, corpus-based linguistics on non-NA anaphora in languages other than English include, e.g., Vieira, Salmon-Alt, Gasperin, Schang, and Othero (2002) on French and Portuguese; Böhmová et al. (2003) on Czech; Navarretta and Olsen (2008) on Danish and Italian; Recasens (2008) on Spanish and Catalan; and Nedoluzhko and Lapshinova-Koltunski (2016) on Czech and German. Needless to say, the processing of non-NA anaphora in languages other than English constitutes an important avenue for future research.

constituents representing non-NAs in the given context (e.g., Strube and Müller 2003; Müller 2008; Chen, Su, and Tan 2010; Kolhatkar, Zinsmeister, and Hirst 2013b; Jauhar et al. 2015; Marasović et al. 2017).

Finally, in Section 6, we conclude with our suggestions and recommendations for future efforts in annotation and resolution of non-NA anaphora, and we present a list of fundamental questions in the field that remain unanswered.

2. The Phenomenon of Non-NA Anaphora

Example (1) demonstrated a simple case of anaphora, in which the antecedent *Maya* is a simple noun phrase. When the antecedent is a non-nominal constituent, as in Example (2), we refer to the relation between an anaphor and its non-NA as non-NA anaphora. Semantically, such non-NAs typically denote complex entities, such as propositions, facts, events, or situations. These entities are complex because they can contain a number of other entities and events, as well as the relationships between them.

There are different terminologies used for the antecedent (i.e., the surface linguistic constituent) and its interpretation. For instance, Byron (2004) uses Luperfoy's (1991) terminology, referring to the antecedent as the *sponsor*, and Byron (2003) uses the term *linguistic anchor* to indicate that these are not ordinary nominal antecedents. The actual meaning of an antecedent is referred to as its **referent** or **interpretation**. In this article, we will use the term **antecedent** to refer to the expression that most closely represents the interpretation of the anaphor, so far as it is overtly realized, and **referent** to refer to the interpretation itself.

A few simple examples of non-NA anaphora are shown in Example (4). Here, the anaphors are: *this*, *this fact*, and *it*; the antecedents are: the sentence *Women are a rarity in mathematics and engineering* in Example (4a), the sentence *They will not win many races* in Example (4b), and the verb phrase *made her butternut squash recipe* in Example (4c); and the referents are: the proposition that women are a rarity in mathematics and engineering in (4a), the fact that those who run with large lateral motion will not win many races in (4b), and the action of making the butternut squash recipe in (4c).

(4) a. **Women are a rarity in mathematics and engineering.** As a female engineering student, I see **this** every day. (NYT[4])

b. Those who run with large lateral motion are not running well; they may be good runners who are very tired, or simply poor runners. **They will not win many races**, but it is far too simplistic to attribute **this fact** to the "extra distance" that must be covered. (NYT)

c. Anna finally **made her butternut squash recipe** this morning. **It** took her twenty minutes.

These examples are relatively simple examples, where the antecedents precede the anaphor, are given explicitly in the text, and are in the close vicinity of the anaphor. None of these circumstances are necessarily the case. In Example (5a), the antecedent follows the anaphor *this*, creating an instance of **cataphora**. In Example (5b), there is no clear syntactic constituent representing the antecedent of the anaphor *this reason*—instead, the

4 Examples with the abbreviation NYT are from the New York Times Corpus (Sandhaus 2008), `https://catalog.ldc.upenn.edu/LDC2008T19`. All URLs in this article have been checked on June 19, 2018.

antecedent is distributed throughout the preceding text. And in Example (5c), the syntactic constituent representing the antecedent is three sentences away from the anaphor and the actual issue here, that is, the referent is *whether* to allow some form of audio-visual coverage of court proceedings or not, which does not occur explicitly in the text.

(5) a. **This**, I now realize, was a very bad idea—**suggesting we do whatever Terry Crews wants for the day**.[5]

b. Because all of us carry some baggage from our past, **I seldom arrive in Paris**, where work takes me four or five times a year, **without some feeling of being an ugly duckling or,** at any rate, **a small-town person**. No doubt it is for **this reason**—I can think of no other—that I stay in the same hotel, in the same room, and consider the area around the Place Vendôme my neighborhood. (NYT)

c. New York is one of only three states that do not **allow some form of audio-visual coverage of court proceedings**. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on **this issue** hurts public confidence in the judiciary … (NYT)

Linguistic accounts of anaphora usually assume that in processing text or utterances, speakers and hearers build a **discourse model** (Kamp 1979; Webber 1979), a mental model of the current discourse state, which is dynamically updated as new utterances are processed. A discourse model contains representations of entities that have been referred to in the discourse up to now, attributes of these entities, and the relationships between them. The entities are called **discourse entities** or **discourse referents** (Karttunen 1976). To determine the referent of a nominal anaphoric expression, a suitable antecedent with the appropriate features, for example, matching gender and number, has to be found, whose discourse referent then serves as the referent of the anaphor. With non-NA anaphoric expressions, on the other hand, there is not necessarily a discourse referent available in the discourse model that the anaphor could refer to. Interpreting non-NA anaphora is therefore often said to involve additional steps of interpretation (Webber 1991) (cf. Section 3.2.1).[6]

2.1 Other Terminologies Describing Similar Phenomena

The phenomenon we discuss in this article has been a subject of interest for linguists, philosophers, and computational linguists for decades. Consequently, it has been addressed in various contexts from a variety of perspectives, as discussed in the following sections.

2.1.1 Abstract Anaphora. Asher (1993), Navarretta (2007), and Dipper et al. (2011) use the terms *abstract anaphora* or *abstract object anaphora*, as in this phenomenon the anaphor refers to an *abstract object*, such as a fact, an event, a proposition, or a situation, in

5 Source: Joel Stein. Crews control. *Time*, September 11, 2014. http://time.com/3326577/terry-crews-wont-hit-the-brakes/.

6 Of course, there are also cases of NA anaphora where no obvious antecedent is available and the anaphor's referent must be inferred, e.g., in the case of bridging relations (Clark 1975).

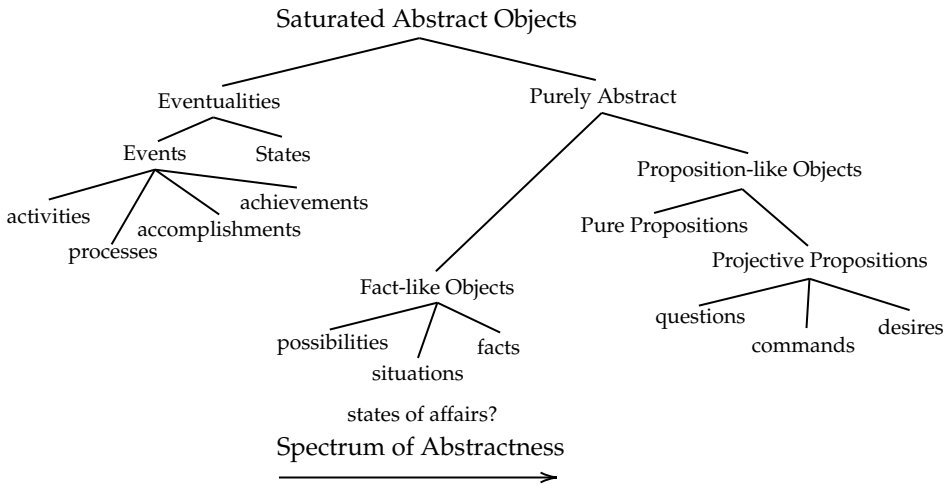


Figure 1
Ashers's typology of saturated abstract objects (Asher 1993, page 57).

contrast to a concrete object, such as a person or a location. Asher formalized the notion of an abstract object by extending Vendler's (1967) approach of using linguistic tests to differentiate various types of abstract objects. The resulting typology (Figure 1) makes a broad distinction between **eventualities** (i.e., events and states, which have spatial, temporal, and causal properties and can be observed by the senses) and **purely abstract objects** (i.e., facts and propositions, which do not have a spatiotemporal location and are not perceivable by the senses but are only mentally conceivable; e.g., Asher 1993, page 57). According to Asher (1993, page 86), eventualities are similar to concrete objects in that they can be directly introduced into the discourse model by some syntactic construction. Whereas concrete objects are introduced by noun phrases (or, more precisely, by their determiners), eventualities are introduced by finite clauses (or, more precisely, by their inflectional marking). In contrast, facts or propositions are introduced by the semantic constraints imposed by specific nouns, such as *fact*, or verbs, such as *believe*, which require their arguments (e.g., a *that* clause) to be of a certain type (Asher 1993, pages 116 and 175).

Asher collectively calls the events, states, processes, propositions, facts, and similar entities that populate these two categories **saturated abstract objects**. They are "saturated" in the Fregean sense that they are themselves either true or false, whereas properties or concepts, although abstract, are only true or false as applied to their arguments (Asher 1993, page 15). It is primarily this category of objects—saturated abstract objects—to which non-NA anaphors refer.

2.1.2 Discourse Deixis. Another popular term is *discourse deixis* (e.g., Webber 1988, 1991; Eckert and Strube 2000; Byron 2004; Recasens 2008).[7] Webber (1988, 1991), attributing the term to Lakoff (1974),[8] calls non-NA anaphors discourse-deictic because the anaphor deictically points to some part of the discourse model from which it gets its reference.

7 The term *deixis* refers to the linguistic phenomenon in which an expression's reference is determined in relation to its extra-linguistic context, e.g., the time (*now*), place (*here*), or participants (*I, you*) of the utterance. Such expressions are called *deictic* (Huddleston and Pullum 2002, page 1451).
8 Lakoff (1974) uses the term in a broader sense, including both NA and non-NA anaphora.

Webber (1991) states that it makes sense to call the phenomenon discourse *deixis* because such relations are usually signaled by deictic expressions, that is, demonstratives *this* and *that*, compared with *it*. Cornish (2007) contrasts deixis with anaphora, describing them as the poles of a scale: Whereas anaphora involves the retrieval of an existing discourse entity from the current model, deixis shifts the focus to a new discourse entity or a new aspect of an existing entity.

The term discourse deixis has also been used in the literature with a different meaning: According to Levinson (1983), discourse deixis occurs when reference is made to the linguistic form of an utterance rather than its referent or when demonstrative expressions refer meta-linguistically to the preceding or following discourse segments (e.g., *this section*, *this chapter*). One can argue that the antecedents in such cases (i.e., *this chapter* and *this section*) are big chunks of text and therefore non-nominal. However, though these are certainly interesting cases, we do not focus on them in this article.

2.1.3 Impure Textual Deixis. Lyons (1977) distinguishes between three different types of entities: First-order entities are physical objects. Second-order entities are events, states of affairs, and processes (Asher's eventualities), which are located in time and involve first-order entities and interactions between them. Third-order entities are propositions and facts (Asher's purely abstract objects), which have no spatiotemporal location, and involve first- and second-order entities and the interactions between them.

For anaphoric relations, Lyons introduced the term *textual deixis*, which describes the deictic relation obtained between a referring expression such as a pronoun and a piece of text. He distinguishes between *pure textual deixis*, where the referring expression refers to a textual unit as such (similar to Levinson's [1983] notion of discourse deixis), and *impure textual deixis*, where the expression is related to the third-order entity denoted by a textual unit, such as a fact or a proposition. If the relation involves a second-order entity (e.g., an event), it is not clear whether Lyons considers this relation an instance of impure textual deixis or simply ordinary anaphora.

2.1.4 Situational Reference. Fraurud (1992) uses the term *situational reference*. She defines situations as entities representing eventualities (e.g., events, processes, and states) and factualities (e.g., facts and propositions). She uses the term *antecedent* for the clause or sentence that provides the anaphor's referent, but often the anaphor refers to a "larger situation"—for example, a whole sequence of events.

2.1.5 Non-nominal Direct and Indirect Anaphora. Gundel, Hedberg, and Zacharski (2004) and Hedberg, Gundel, and Zacharski (2007) use the terms *non-nominal direct* and *indirect anaphora*. They operationalize this terminology as follows. An anaphoric relation is *direct* if the anaphor's referent is the same as the antecedent's referent, and it is *indirect* if the interpretation of the anaphor depends on that of the antecedent but they are not coreferential because the interpretation involves an additional step. Example (6), from Hedberg, Gundel, and Zacharski (2007), is an example of a direct anaphoric relation because both the anaphor and the antecedent refer to the event of the stock doubling on its first day of trading. In contrast, in Example (7), from Hedberg, Gundel, and Zacharski (2007), the clausal antecedent introduces the *state* of Giuliani being sleepy and the marked anaphor refers to the *fact* that he was sleepy, so the anaphor is not coreferential with the antecedent here, and it would be classified as an instance of indirect anaphora.

(6) The winner was Internet Capital Group, a company that invests in other Internet companies. **It more than doubled its first day of trading**, Aug. 5., and **that** was just the beginning.

(7) Mayor Rudolph Giuliani, who gave himself the job of ubiquitous master of ceremonies of the city's New Year celebration, said he began his last day of the 1900s at 5:30 a.m. having trouble getting his lights on. "I was convinced that it was Y2K," the mayor said, but **"actually I was sleepy." This** perhaps explains an interesting mishap…

Botley (2006) provides the following characteristic properties for indirect anaphora: (a) The antecedent is not nominal and is difficult to define directly, (b) the link between anaphor and antecedent is not one of coreference, and (c) the hearer may have to carry out a complex process of inference to arrive at the antecedent. Botley considers three main types of indirect anaphora: textual deixis[9] (Lyons 1977), situational reference (Fraurud 1992), and labeling (Francis 1986). We discussed textual deixis and situational reference in the previous subsections, and we will discuss labeling (i.e., shell nouns) in Section 3.1.

2.1.6 Complex Anaphora. Consten, Knees, and Schwarz-Friesel (2007) coin the term *complex anaphora*, where anaphors are nominal expressions referring to propositionally structured referents, such as propositions, states, facts, and events. They define two criteria for complex anaphora: First, the antecedent has to be a syntactically complex entity—it must consist of at least one clause; and second, the antecedent must denote a conceptually complex item.[10] Consten, Knees, and Schwarz-Friesel define a conceptually complex item as a second- or third-order entity, according to Lyons' (1977) hierarchy (see Section 2.1.3).

2.1.7 Extended Reference and Text Reference. Halliday and Hasan (1976, pages 52–53, 66–70) distinguish between two kinds of references of demonstrative pronouns and the pronoun *it*: *extended reference* and *text reference*.[11] An example from Halliday and Hasan (1976, page 52) is given in Example (8).[12] The first instance of *it* in the example refers to *curtseying while you're thinking what to say*, which they call *extended reference*, as the reference is no longer to a person or object but to a whole process or complex phenomenon, and the referent is expressed by a clause or string of clauses instead of a simple noun phrase. In contrast, the second instance of *it* is a case of text reference because it requires its referent to be transmuted into the *fact* that *curtseying while you're thinking what to say saves time*.

(8) [The Queen said:] 'Curtsey while you're thinking what to say. **It** saves time.' Alice wondered a little at this, but she was too much in awe of the Queen to disbelieve **it**.

9 Botley calls this type text/discourse deixis. Discourse deixis is here understood in the sense of Levinson (1983), which is very close to Lyons' pure textual deixis.

10 Both conditions are necessary to distinguish non-NA anaphora from bridging relations (see footnote 6): Example (ia) is a case of a non-NA anaphor with *this incident* referring to the biting event reported in the previous sentence. Example (ib) is a case of a bridging relation: The expression *the scars* does not refer to an event but to a concrete entity, which is inferred from an entity involved in the biting event. (We owe this example to Manfred Consten, personal communication.)

(i) a. **One year ago a dog bit me**. **This incident** traumatized me.
b. One year ago a dog bit me. **The scars** are still visible today.

11 On page 66, Halliday and Hasan (1976) use the term *reference to 'fact'* for text reference.

12 The example is originally from Lewis Carroll's *Through the Looking-Glass* (1871).

Table 1
Overview of terminology used for non-NA anaphora.

I. Approaches that make a distinction between eventualities and factualities		
	Reference to eventualities	**Reference to factualities**
Halliday and Hasan (1976)	extended reference	text reference
Hedberg, Gundel, and Zacharski (2007)	direct anaphora	direct + indirect anaphora
Lyons (1977)		impure textual deixis

II. Approaches that do not make a distinction between eventualities and factualities	
	Reference to abstract objects
Asher (1993) and others	abstract anaphora
Webber (1988) and others	discourse deixis
Fraurud (1992)	situational reference
Consten, Knees, and Schwarz-Friesel (2007)	complex anaphora

2.1.8 Comparison of Terminology. The overview of the terminology in the previous sections showed that the approaches have a lot in common. In particular, many approaches distinguish between two subclasses of abstract entities:

1. Events and states: These are called *eventualities* by Asher and Fraurud, and *second-order entities* by Lyons.
2. Facts and propositions: These are called *purely abstract objects* by Asher, *factualities* by Fraurud, and *third-order entities* by Lyons.

The two subclasses make up a superclass, called *saturated abstract objects*, by Asher and *situations* by Fraurud. In this article, we will use the terms **eventualities** and **factualities** for the two subclasses, and **abstract objects** or **abstract entities** for the superclass (contrasting these with **concrete objects/entities**).

The different kinds of terminology introduced here can roughly be divided in two classes, depending on whether they make a (terminological) distinction between anaphoric relations involving eventualities and relations involving factualities or not, as shown in Table 1.[13]

Another distinction is whether non-NA anaphora is seen as a type of anaphora or deixis. If it is considered a type of anaphora, finding the referent involves first determining the antecedent. The anaphor's referent is then either the same as the antecedent's referent (i.e., they are coreferent), or it is derived from it. If non-NA anaphora is viewed as an instance of deixis, no antecedent is involved; rather, the anaphor's referent is determined by pointing to a region of the discourse or discourse model. As already stated, in this article we use the term antecedent to refer to the expression that most closely represents the interpretation of the anaphor, so far as it is overtly realized. We also occasionally say that an anaphor refers to a non-NA, meaning that the anaphor refers to some abstract referent that is represented in the text by the non-NA.

13 The division made by Hedberg, Gundel, and Zacharski (2007) cuts across these two classes, in a way, because they consider all instances that do not require coercion (see Section 3.2.1) to be direct. However, all anaphoric relations involving eventualities are instances of direct anaphora.

Table 2
Typology of anaphora on syntactic and semantic scales. Non-NA anaphora as discussed in this article falls in Cell D.

		Semantics	
		Concrete objects	*Abstract objects*
Syntax	*Nominal*	A	B
of antecedent	*Non-nominal*	C	D

2.2 Typology of Anaphora on Syntactic and Semantic Scales

The overview of different terminologies showed that most of them highlight semantic aspects of the phenomenon. In the context of building computational approaches, we argue for the use of the syntactic notion of non-NA anaphora.

Table 2 describes the phenomenon of anaphora in general from syntactic and semantic aspects. On the syntactic scale, we show two ends of the scale: non-nominal and nominal.[14] On the semantic scale, we have abstract objects on the one hand, which include events, states, and processes along with purely abstract objects, such as propositions and facts; and on the other hand we have concrete objects. Cell A represents concrete nominal objects (cf. Example (1)). Most of the coreference and anaphoric relations annotated in corpora such as MUC-7[15] and OntoNotes (Weischedel et al. 2013) fall in this category. OntoNotes also includes annotated instances of certain types of relations involving eventualities, so that OntoNotes also covers a subset of Cell D relations (but see Section 4.2.5).

Cell B is representative of examples such as Example (9), where the surface linguistic form of the antecedent *same-sex marriage* is nominal but it represents an abstract concept. Some approaches we discuss in this article do include such examples in their annotation and resolution (e.g., Kolhatkar 2015). Note that one can argue that the actual issue here is not just the concept of same-sex marriage but rather whether to allow same-sex marriage or not, which is only implicitly stated in the text.

(9) **Same-sex marriage** is currently one of the most divisive political issues in our nation. Analyzing **this issue** will help us understand what is happening in our country, and where we might go from here.[16]

Cell C represents examples having non-NAs but that do not represent abstract objects. Metalinguistic anaphoric expressions such as *this section* or *this statement*, which refer to the spatiotemporal coordinates of the text or act of utterance, are examples of this category. The antecedents of such expressions are clearly non-nominal but they do not usually represent abstract objects according to our definition. Finally, Cell D is

14 Note that the table only shows two ends of the syntactic spectrum and the boundary between them is fuzzy. For instance, gerunds denoting events, such as the phrase *the mayor's throwing of the pizza in the guest of honor's face* (Asher 1993, page 16), fall somewhere in between the two.

15 `https://catalog.ldc.upenn.edu/LDC2001T02`.

16 Adapted from Martha Nussbaum. A right to marry? Same-sex marriage and constitutional law. *Dissent*, Summer 2009. `https://www.dissentmagazine.org/article/a-right-to-marry-same-sex-marriage-and-constitutional-law`.

representative of the Examples (4) through (8). We consider members of this category instances of non-NA anaphora.[17]

2.3 Summary and Outlook

In this section, we presented some simple and some complex examples of non-NA anaphora. We saw that different terminologies from the literature highlight slightly different aspects of the phenomenon. They can be divided into two classes, depending on whether they make a distinction between eventualities and purely abstract objects or not. We discussed two important aspects of non-NA anaphora: One concerns its syntax and the other its semantics. The syntactic aspect is about identifying the surface linguistic constituent, which is referred to as an antecedent, and the semantic aspect is related to identifying the actual referent, interpretation, or meaning of the anaphor.

We define non-NA anaphora as a syntactic notion, that is, in terms of the syntactic shape of the antecedent, and show where the phenomena we discuss in this article fall on the semantic and syntactic scales. We choose to use a syntactic notion of non-NA anaphora rather than one of the existing terminologies for two reasons. First, having a non-nominal syntactic type is a characteristic property of proposition-like entities (e.g., Passonneau 1989; Gundel, Hedberg, and Zacharski 1993). And second, we believe that from a computational linguistics perspective, whether the entity to which an antecedent refers is abstract or concrete, or whether it is indirect reference or not, does not in and of itself present a particular challenge to automatic resolution. It is the recovery of non-NAs that presents a distinctive and challenging problem.

In the next section we will see that the syntactic and semantic aspects of the phenomenon are not necessarily independent. For instance, facts are more likely to be represented with *that* clauses than with gerund phrases.

3. Linguistic Properties of Non-NA Anaphora

This section is divided into two main parts, examining the range of linguistic forms with which non-NA anaphors are realized in English, and discussing the properties of such anaphors and their antecedents from a linguistic perspective.

3.1 Realization of Non-NA Anaphors

3.1.1 Pronouns. A very common way to signal non-NA anaphora is with the singular demonstrative pronouns *this* and *that*. Example (10) shows *this* and *that* referring to non-NAs. In Example (10a), *this* refers to the fact that my daughter is all grown up and going away to college; and in Example (10b), *that* refers to the proposition that this policy will help the poor.[18]

(10) a. **My daughter is all grown up and going away to college. <u>This</u>** is so bittersweet.

b. He says **that this policy will help the poor**. But we all know **<u>that</u>**'s not true.

17 Note that some approaches, e.g., Schiffman (1985) (cf. Section 4.2.1), consider clauses introduced by *that* or *whether* to be nominals.

18 Note that plural demonstrative pronouns *these* and *those* are not used to refer to non-NAs. For instance, when we want to refer to multiple events or facts we are likely to use plural forms, either in a copular construction (*these are the facts*) or as a noun phrase (*those events*). But referring to multiple facts or events with just *these* and *those* is not possible; see Halliday and Hasan (1976, page 66).

The personal pronoun *it* may also be used to refer to non-NAs. In Example (11), the pronoun *it* refers to the fact that John is a gifted player.

(11) **John is a gifted player**, and he knows **it**.

Although the pronouns *this*, *that*, and *it* have the capacity to refer to non-NAs, there is strong evidence showing that demonstrative pronouns are more likely. Two kinds of preferences with regard to pronominal anaphors can be discerned.

First, we can examine whether, given a certain type of antecedent, a certain anaphor is preferred. There are clear correlations between the syntactic type of the antecedent and the choice of pronominal anaphors: Nominal antecedents are associated with *it* and non-NAs with demonstrative pronouns. In a corpus of career-counseling interviews, Schiffman (1985, Section 5.5.2) found that from a total of 298 non-NAs, *that* is used in 78.2%, and *it* in 21.8% (there were only very few instances of *this* in the corpus); if the antecedent is a sentence or paragraph, *that* is used in 88.9% of these cases. With nominal antecedents, her results do not show a clear preference: From a total of 227 nominal antecedents, 51.5% are referred to by *it* and 48.5% by *that*.[19] In a corpus of six texts from different domains, Webber (1991) observed that 98% of 81 nominal antecedents occurred with *it* and only 2% with *this/that*. There were 96 cases of clausal antecedents, and *it* was used in 16%, *this* in 65%, and *that* in 20%.

The second preference is in the opposite direction: Given a certain type of anaphor, we ask which type of antecedent is more likely. There are also correlations between the type of the pronoun and the syntactic type of the antecedent it refers to. For example, the majority of instances of *it* refer to nominal, non-abstract objects, as in Example (12).

(12) There was **a black kitten** in the backyard. **It** was chasing a squirrel.

The figures reported by Webber (1991) show that 84% of (referential) *it* instances refer to nominal antecedents and 16% to clausal ones. In contrast, 98% of *this/that* instances refer to non-NAs. In a corpus of dialogues, Byron (2004, page 10) finds that from a total of 260 pronouns, 74% of the personal pronouns (including *it*) refer to a nominal antecedent but only 23% of the demonstrative pronouns do, whereas only 7% of the personal pronouns refer to a non-NA, in contrast to 32% of the demonstrative pronouns. Halliday and Hasan (1976, page 66), who considered demonstratives in general (both pronouns and determiners), counted 51 demonstratives in total in the last two chapters of *Alice's Adventures in Wonderland*: 43% *this*, 47% *that*, 6% *these*, and 4% *those*. Of these instances, 61% were used with non-NAs. In their analysis of instances of *it*, *this*, and *that* in the Santa Barbara corpus of spoken American English Part-I (Du Bois et al. 2000–2005),[20] Gundel, Hedberg, and Zacharski (2004) observed that among 56 examples of demonstrative pronouns, 24 (42.9%) were classified as non-NA anaphors and 21 (37.5%) were classified as NA anaphors. In contrast, among 2,046 instances of third person pronouns, only 110 instances (5.38%) had non-nominal antecedents.

3.1.2 Shell Nouns. Another common way to signal non-NA anaphora is with **shell nouns** (Schmid 2000; Kolhatkar 2015). These are abstract nouns, such as *fact*, *issue*, or

19 These figures are not directly specified in Schiffman (1985). We consider only Schiffman's "true NPs" to be nominal antecedents and her "sentence-like NPs" to be non-nominal antecedents (see also Section 4.2.1). Details on how these figures (and others in this survey) have been calculated can be found at `https://github.com/kvarada/non-NA_Resources`.

20 `http://www.linguistics.ucsb.edu/research/santa-barbara-corpus`.

Table 3
Lexico-grammatical patterns of shell nouns from Schmid (2002, adapted from page 24). The patterns are marked in italics. Shell noun phrases are underlined, and the antecedents or structurally related phrases providing the shell content are marked in bold, as usual. '*' marks additional patterns discussed by Schmid. All examples are from the New York Times Corpus (Sandhaus 2008).

	Anaphoric relations	
1	*th-N*	**Living expenses are much lower in rural India than in New York**, but ***this fact*** is not fully captured if prices are converted with currency exchange rates.
2	*th-be-N*	**People change**. *This is* ***a fact***.
	Structurally-determined relations	
3	*N-be-to*	***Our plan*** *is* ***to*** **hire and retain the best managers we can**.
4	*N-be-that*	***The major reason*** *is* ***that*** **doctors are uncomfortable with uncertainty.**
5	*N-be-wh*	Of course, ***the central, and probably insoluble, issue*** *is* ***whether*** **animal testing is cruel**.
6	*Sub-be-N* *	If the money is available, however, ***cutting the sales tax*** *is* ***a good idea***.
7	*N-to*	***The decision*** ***to*** **disconnect the ventilator** came after doctors found no brain activity.
8	*N-that*	Mr. Shoval left open ***the possibility*** ***that*** **Israel would move into other West Bank cities**.
9	*N-wh*	If there ever is ***any doubt*** ***whether*** **a plant is a poppy or not**, break off a stem and squeeze it.
10	*N-of* *	***The concept*** *of* **having an outsider as Prime Minister** is outdated.

decision, that have the capability to encapsulate and refer to propositional content. These nouns are known by a great variety of names in the literature, including *container nouns* (Vendler 1968), *type-3 vocabulary* (Winter 1977), *anaphoric nouns* (Francis 1986), *label nouns* (Francis 1994), and *carrier nouns* (Ivanič 1991). They are likewise included in Halliday and Hasan's (1976) concepts of *extended reference* and *text reference*. In this article, we use Schmid's (2000) term *shell nouns*, which derives from these nouns' tendency to function as conceptual shells for propositional content.

To be a shell noun is not an inherent property of nouns themselves; rather, it is a property of particular instances of these nouns, which can be characterized individually as shell noun usages. In the context of shell nouns, the term **shell content** is often used to refer to the text that provides the interpretation of the shell noun phrase.[21] Schmid (2000) observed a number of lexico-grammatical patterns in which shell nouns tend to occur. Table 3 shows these patterns. Shell nouns may refer anaphorically to their shell content, as shown on lines 1 and 2,[22] or they can be structurally related to their shell content, as shown on lines 3 to 10. These relations include copula structures (lines 3–6)[23] and postnominal complement and modifier clauses (lines 7–10). As the pattern labels

21 The concept of shell content is similar to the concept of an antecedent except that in some shell noun constructions, the term antecedent is not quite appropriate. That said, to be consistent, we use the term antecedent for shell content except when it is absolutely necessary to use the term shell content.

22 Schmid (2000) clarifies that "[i]n a way, the pattern *th-be-N* is a blend of the copular type *N-be-cl* and the anaphoric type *th-N*." (page 25).

23 The pattern *Sub-be-N* is not part of Schmid's (2000) original list but he discusses it in his example (3.5') on page 26.

N vs. *th-N* suggest, in structurally determined shell noun constructions, definite noun phrases and even indefinite uses are common, whereas in anaphoric constructions, the noun is most frequently used with a demonstrative determiner.

Note that although plural pronouns are only very rarely used in non-NA anaphora (if at all), plural demonstrative noun phrases as in *th-N* (Example (13)) and *th-be-N* (Example (14)) are possible.

(13) Before applying for an appropriate job, you need to think about **which city you want to live in, whether you want an academic or an industry position, whether you are willing to live away from your family**. The application process is smooth if **these decisions** are made in advance.

(14) Today we hear **heated rhetoric between nuclear powers**; we see **a deeply divided United States**; we watch as **populist nationalisms take hold around the world**. **These** are some of **the major global obstacles** that will trouble any engaged citizen of our planet.

Shell nouns occur frequently in many varieties of text, but they are especially frequent in argumentative texts, such as political debates, news articles, and academic discourse (Schmid 2000; Botley 2006; Kolhatkar 2015). Schmid (2000) provides a list of 670 nouns that can be used as shell nouns and occur in the patterns provided in Table 3.[24] He observed that shell nouns such as *fact*, *idea*, *point*, and *problem* are among the one hundred most frequently occurring nouns in a corpus of 225 million running words of British English.[25] Because of the detailed existing documentation on shell nouns, their lexico-grammatical patterns, and the relatively stable word order of English, it is possible to gather shell noun instances largely automatically, which makes them, from a computational linguistics perspective, a great source of information for studying the phenomenon of non-NA anaphora.

3.1.3 Reflexives, Pro-verbs, Pro-actions, Pro-adjectives, Adverbs, etc. Apart from the main constructions described earlier, there are a number of other constructions that may also be used in non-NA anaphora.

The first one is reflexive pronouns. Non-NA anaphora can be realized with the reflexive pronoun *itself*, as shown in Example (15).

(15) **That we are hungry** shows **itself** in our crankiness.

The second one is the *pro-verb* construction where the anaphor is a form of *do* (cf. Halliday and Hasan 1976, pages 125–126; Hirst 1981, page 19; Miller 2011).[26] The following example from Hirst (1981, page 19) contains such a construction. Here, the pro-verb *does* refers to the verb phrase *orders sweet and sour fried short soup* as its antecedent.

(16) When Ross **orders sweet and sour fried short soup**, Nadia **does** too.

Then there are pro-action constructions in which *do* is used in conjunction with *so*, *it*, or demonstrative pronouns (Hirst 1981; Miller 2011) and whose antecedents usually denote actions. Example (17), from Hirst (1981, page 20), demonstrates such a usage. Here, the anaphor *does it* refers not to the previous event but to the action—Sue cooks Ross's dinner and not her own.

24 Note that Schmid's lists of shell nouns and their patterns are not exhaustive.

25 Schmid used the Bank of English corpus, which is jointly owned by HarperCollins Publishers and the University of Birmingham `http://www.titania.bham.ac.uk/docs/`.

26 Note that these pro-action constructions may also be considered cases of verb phrase ellipsis.

(17) Ross makes his dinner on weekdays, but when she stays the weekend Sue **does it** for him.

Cornish (1992) points out non-NA anaphora realized with the adverb *so*. In Example (18), the antecedent is an adjective phrase, and in Example (19), the antecedent is a prepositional phrase (both examples are from Cornish 1992).

(18) I'm **extremely busy** at the moment, and expect to be **so** for the next two hours at least.

(19) The entire factory is **on strike**, and is forecast to stay **so** for some considerable time.

These constructions have not yet received much attention in computational approaches, and we do not discuss them in the rest of the article.[27]

3.1.4 Summary of Expressions Used as Anaphors Referring to Non-Nominal Antecedents. As we saw in the previous subsections, non-NA anaphors can be realized with a variety of expressions. We will use the following terms throughout this article.[28]

1. Personal pronoun: *it*
2. Demonstrative pronouns: *this*, *that*
3. Shell nouns: Any of a number of nouns that may occur in one of the patterns listed in Table 3
4. *This* NP: The subset of shell nouns that occur with a demonstrative determiner, e.g., *this issue, these facts, that situation*
5. Demonstratives: an umbrella term for demonstrative pronouns and *this* NPs

Note that although the expressions in this list have the capacity to refer to non-NAs, not all instances of these expressions refer to such antecedents. For instance, we have already seen examples of the pronoun *it* referring to nominal antecedents, as in Example (12). Other instances of *it* are pleonastic, namely, they do not refer to any specific entity but serve as syntactic placeholders, as in Example (20).

(20) a. **It** is gorgeous out with blue skies and big fluffy white clouds and a light breeze.

b. **It** is snowing.

From a computational perspective, automatically identifying whether a given instance of a shell noun or *this*, *that*, or *it* refers to a non-NA or not is not possible with regular expression-based queries alone. At a minimum, we need part-of-speech

27 Anand and Hardt (2016) present a machine learning approach to sluicing, an elliptical phenomenon closely related to the verb phrase ellipsis in Example (16).

28 Note that the terms pronoun and demonstrative are sometimes used differently in the literature. We reserve the term pronoun for cases where the pronoun substitutes an entire NP, e.g., as in Example (ia). If it is used like an article, we call it a determiner; see Example (ib). Other work would call the expression *this* a demonstrative pronoun in both uses (e.g., Artstein and Poesio 2006).

(i) a. I know **this**. (pronoun)

b. I know **this issue**. (determiner)

The term demonstrative is sometimes used to refer to demonstrative pronouns only (as in Example (ia)), excluding full NPs with a demonstrative determiner (as in Example (ib)), e.g., by Eckert and Strube (2000).

information to identify lexico-syntactic patterns from Table 3 or to weed out instances of relative pronouns—for instance, as shown in Example (21).

(21) The police said the accident **that** happened last night was unavoidable.

A few computational approaches have been proposed for detecting instances of non-referential pronouns (e.g., Boyd, Gegg-Harrison, and Byron 2005; Müller 2006; Bergsma and Yarowsky 2011; also cf. Uryupina, Kabadjov, and Poesio 2016). That said, this problem remains unsolved. For instance, Loáiciga, Guillou, and Hardmeier (2017) recently showed that distinguishing different usages of *it*, in particular between nominal anaphoric use and reference to events, remains a complex task.

3.2 Properties of Non-NA Anaphors and Antecedents

Non-NA anaphora demonstrates an interplay between different linguistic levels, and in this section, we will discuss the constraints and preferences present at each level.

3.2.1 Lexical and Semantic Preferences. Non-NA anaphora is in many crucial respects a semantic phenomenon and a number of semantic preferences and properties are associated with it. These properties can be broadly categorized into three types: those imposed by the context of the anaphor, by the anaphor or the antecedent itself, or by the context of the antecedent.

Preferences imposed by the anaphor context. It is often assumed that a non-NA does not introduce various types of semantic objects into the discourse model but, instead, that one semantic object—for example, an event-type discourse referent—is introduced, which can be transformed into other types when it is referred to anaphorically. Webber (1988), Eckert and Strube (2000), and Byron (2004) refer to this property as *referent coercion* and Webber (1991) calls it *ostension*. An example of referent coercion adapted from Eckert and Strube (2000) is shown in Example (22).

(22) **John crashed the car.**
   a. His girlfriend couldn't believe **it**. (proposition)
   b. **It** happened yesterday at 10 in the morning. (event)
   c. **This** shows how careless he is. (fact)

Here, the antecedent *John crashed the car* denotes an event, whereas the pronouns refer to a variety of semantic types, depending upon the context. For instance, in Example (22b), *this* refers to the event of crashing, and, as in Example (22c), *this* refers to the fact that John crashed the car.

The predication context of an anaphoric expression plays an important role in identifying the **semantic type** of the referent. For our purposes, semantic type is an abstract object, as defined by Asher (1993) (e.g., fact, proposition, or event). For instance, the verb *believe* can only be applied to an object argument that represents a proposition (e.g., see Example (22a)), and *happen* can only be used with subjects denoting some sort of event (Asher 1993, pages 22, 192) (e.g., see Example (22b)).

The anaphor's semantic type preferences can constrain the syntactic type of antecedents. In Example (23), the anaphor refers to a concept and, accordingly, the antecedent is realized by a verb phrase (VP) instead of a sentence.

(23) John **crashed the car.** Jane did **that** too. (concept)

Eckert and Strube (2000) introduce the concepts of A(bstract)-incompatibility and I(ndividual)-incompatibility. A particular instance of an anaphor is described as I-incompatible if, given its context, it cannot refer to an individual, concrete entity. Typical contexts for I-incompatible anaphors include those in which adjectives are used that can only be applied to abstract entities, such as in *x is true* or *x is correct*, and those in which the anaphor is equated with abstract entities, such as facts or reasons: *x is why he's late* implies that *x* is a reason, an abstract entity. A-incompatibility refers to the opposite situation, in which the context precludes the anaphor from referring to abstract objects, as in *x is loud* or *eat/drink/smell x*.

Preferences imposed by the anaphor or the antecedent. According to Hegarty, Gundel, and Borthen (2001), the degree of world immanence of an entity and the degree of its individuation contribute in deciding whether it can be referred to by demonstratives or the pronoun *it*. As Asher (1993) points out, eventualities have causal, spatial, and temporal properties, and thus they have high world immanence. The personal pronoun *it* preferably refers to such entities. In contrast, factualities such as facts, situations, or propositions have very low world immanence, and demonstratives are highly preferred to refer to them; see Example (24) from Gundel, Hedberg, and Zacharski (2005). Example (24a) shows that *it* (and *that*) can refer to the event of insulting. If one wants to refer to the situation, demonstrative pronouns are clearly preferred over the personal pronoun *it*; compare Examples (24b) and (24c).

(24) **John insulted the ambassador.**
   a. **It/that** happened at noon. (event)
   b. **That/this** was intolerable to the embassy. (situation)
   c. ?? **It** was intolerable to the embassy. (situation)

Another preference imposed by an antecedent is with respect to tense. Schmid (2000, pages 104–105) notes that there is a strong present or past tense preference (as opposed to future tense or modal forms) in antecedents whose semantic type is *fact*, which he relates to the semantics of facts: Future facts are not knowable and therefore speakers will prefer other shell nouns for these types of content.

Preferences imposed by the antecedent context. Hegarty, Gundel, and Borthen (2001) describe how the complements of *bridge verbs*, such as *think*, *believe*, and *say*, are typically accessible to reference with demonstratives, but not with the personal pronoun *it*, as shown in Example (25), from their paper.

(25) A: Alex believes that **the company destroyed the file**.
   a. B: **That**'s false; the file has been submitted to the district judge.
   b. B: # **It**'s false; the file has been submitted to the district judge.

On the other hand, entities introduced with complements of *factive verbs*, such as *verify* or *know*, are equally accessible to demonstratives and the pronoun *it*, as shown in Example (26), from their paper. According to Hegarty, Gundel, and Borthen (2001), factive verbs mark the entity expressed by the complement clause as already familiar, so that successive reference by *it* is possible.

(26) A: Alex verified that **the company destroyed the file**.
   a. B: **That**'s false; the file has been submitted to the district judge.
   b. B′: **It**'s false; the file has been submitted to the district judge.

3.2.2 Syntactic Preferences. According to Asher (1993, page 226), the range of syntactic constructs of abstract antecedents is quite broad. Some examples of the different linguistic constructions that may function as abstract antecedents are:

1. *That* clauses; e.g., *John believed* ***that Mary was sick***
2. Infinitival phrases; e.g., *Fred wanted* ***to go to the movies***
3. Naked infinitive complements; e.g., *John saw* ***Mary arrive***
4. Noun phrases that appear to denote proposition-like entities; e.g., ***The claim that Susan got a C on the test*** *was surprising*

Moreover, it seems evident that the semantic type of the antecedent also suggests corresponding syntactic realizations. Schmid (2000, page 381) provides the frequencies of 670 shell nouns from the 225 million-word corpus of the British section of the Bank of English corpus. For each shell noun, he provides the frequency distribution of that shell noun across the different lexico-grammatical patterns from Table 3. Among other tendencies, it is evident from these frequencies that *purposes* and *decisions* are more likely to be represented by *to*-clauses; *explanations* and *facts* by *that*-clauses; and *questions* and *issues* by *wh*-question clauses.

Passonneau (1989) analyzed local contexts of *it* and *that* in four career counseling interviews. She observed that grammatical functions (e.g., subject) play an important role: If both the anaphor and its antecedent are subjects, *it* is far more likely than *that*; if either one or both are not subjects, *that* is more likely. Moreover, the syntactic type of the antecedent is also relevant: If the antecedent is pronominal, *it* is more likely than *that*. If the antecedent is non-nominal or a gerund, *that* is more likely. With canonical NP antecedents, both are equally likely.

3.2.3 Discourse-Level Preferences. In the literature, discourse-level properties of non-NA anaphora are discussed in terms of three notions: **salience**, **focus**, and **topic**. These notions are usually grounded in **Centering Theory** (Grosz and Sidner 1986) and theories explaining the cognitive status of these expressions.

Centering Theory models coherence and salience. It assumes that each *utterance* introduces new discourse entities into the discourse, which are organized in a **focus space**. Focus spaces, which constitute the **global focus**, are ordered in a stack so that only entities of the most recent space are in the **local focus** and, e.g., accessible for subsequent reference by pronouns. The discourse entities introduced by an utterance are *ranked* and the most highly ranked entity is referred to as the **preferred center (CP)**. The **backward-looking center (CB)** of an utterance is defined as the highest ranked element of the previous utterance that is realized in the current utterance. The theory itself keeps many notions, such as utterance or ranking, deliberately open, and researchers define these notions differently, depending on their theory and the language under investigation. An utterance is generally defined as a sentence or a clause. Ranking is generally based on the grammatical function (e.g., subject is ranked higher than object) or on information status (e.g., hearer-old entities are ranked higher than hearer-new entities) (cf. Poesio et al. 2004).

The CP of an utterance is considered the most salient entity. The notion of CB is the closest concept to the traditional notion of topic (e.g., Taboada and Wiesemann 2010). Another relevant term is **activation**, which can be defined in the framework of Centering Theory in different ways. A discourse referent can be considered activated if it is in the local focus, or in the global focus and sufficiently salient (Poesio and Modjeska 2005).

in focus	> activated	> familiar	> uniquely identifiable	> referential	> type identifiable
{*it*}	{*that*, *this*, *this* N}	{*that* N}	{*the* N}	{indefinite *this* N}	{*a* N}

Figure 2
Givenness Hierarchy from Gundel, Hedberg, and Zacharski (1993, page 275)

The general idea behind the cognitive status of an anaphoric expression is as follows. In language, we use different expressions to refer to the same thing. For instance, a particular fact can be referred to as *a fact*, *the fact*, *this fact*, *that fact*, *this*, *that*, or *it*. The question is, when do writers use a particular anaphoric expression, and what enables readers to interpret this expression appropriately? Many researchers have made claims about the cognitive status of the antecedents of different expressions, and they support their claims with appropriately annotated data.

Gundel, Hedberg, and Zacharski (1993) propose a **Givenness Hierarchy** of six **cognitive statuses** of discourse referents, which reflect a speaker's assumptions about the addressee's knowledge and current state of attention. These statuses determine the necessary and sufficient conditions on the use of each referring form in discourse (cf. Figure 2). For instance, by using *it* as an anaphor, the speaker signals that the expression refers to an entity in focus, whereas with a form such as *this N*, the speaker refers to an activated entity. Gundel, Hedberg, and Zacharski note that pronominal anaphors, as a universal tendency, prefer their referent to at least be activated, which makes sense because the pronouns only have minimal descriptive content, so in order to facilitate identification of their referents they have to be at least activated.

Based on these notions, the following observations have been made in the literature.[29]

Referring to the entities in focus. It has been demonstrated by different researchers in different domains that the pronoun *it* requires its referent to be in an addressee's *focus of attention*, whereas demonstratives only require them to be activated, namely, present in working memory (the global focus) but not necessarily in focus (i.e., in the local focus).

Hegarty, Gundel, and Borthen (2001) start from the hypothesis that the Givenness Hierarchy can explain the findings of Schiffman (1985) and Webber (1991) that nominal anaphora are preferably realized with *it* and non-NA anaphora with demonstratives (see Section 3.1.1). These findings can be explained if entities introduced by nominals are more easily brought into focus than their non-nominal counterparts. They link an entity's property of being in focus with its degree of world immanence (compare Section 2.1). Accordingly, concrete objects are often brought into focus, eventualities less so, and factualities only rarely. Eventualities are more accessible than factualities because they can be directly introduced by clauses, whereas factualities have to be derived from them (see Section 2.1). The findings from their corpus support this: Gundel,

29 Researchers have examined the Givenness Hierarchy in a range of studies. Many of these studies focus on anaphora with nominal antecedents, but here we only discuss studies that consider non-NAs of *it* and demonstratives.

Hedberg, and Zacharski (2002) observed that of 2,046 instances of third-person personal pronouns (including *she, they*, etc.) from the Santa Barbara Corpus of Spoken American English, 83.34% had nominal antecedents and 5.38% were instances of the anaphor *it* with non-NAs. Among these instances, only 14.54% involved facts or propositions, 57.27% involved situations, and 27.27% eventualities. Gundel, Hedberg, and Zacharski classify situations as less abstract than facts and propositions and note that the distinction between eventualities and situations is not always clear.

Hedberg, Gundel, and Zacharski (2007) analyze 321 instances of pronominal *this* (44%) and *that* (56%) in a corpus composed of two issues of the *New York Times*. They define an entity as activated if it is in the local focus but not salient (i.e., it has been introduced in the previous sentence but not in a syntactically prominent position). It is not clear from their paper how many instances are non-NA anaphora but the majority probably are. They observe that of 256 instances for which the annotators were in agreement, 96% of the demonstrative pronouns refer to activated entities that are not in focus. They also find that the majority of these cases are indirect anaphora, requiring coercion. An annotated example from Hedberg, Gundel, and Zacharski (page 35) is shown in Example (27).

(27) <P num="405"> With the exception of Japanese equities, <1> **Americans have been selling more foreign stocks than they have been buying in recent months.** </1> </P>
<P num="406"> But <that ACTIVATED INDIRECT "the situation that Americans have been selling foreign stocks more than buying them" 405.1 A3 num="06"> **that** </that> could change. </P>

In the example, the antecedent is located in the sentence identified as 405, and is marked by the SGML tags <1>…</1>. The tag containing the anaphor *that* carries all annotated features, for example, ACTIVATED for the cognitive status, INDIRECT for the type of anaphoric relation, a description of the anaphor's referent ("the situation that …"), and a pointer to the antecedent (405.1).

Referring to salient (highly ranked) entities. If being in (local) focus is defined via the focus space, all entities introduced in the same utterance share the same degree of being in focus. Such entities, however, can be differentiated by another property, salience. An entity is made salient if it is introduced in a syntactically prominent position (e.g., as the subject or object) or if it is mentioned repeatedly (Gundel, Hedberg, and Zacharski 1993; Gundel, Hegarty, and Borthen 2003). Gundel and colleagues have shown that (unstressed) personal pronouns in general are used to refer to highly ranked, salient entities in discourse, operationalized as the preferred center in Centering Theory. In contrast, demonstratives are associated with focus shift (Gundel, Hedberg, and Zacharski 1988), when the anaphor does not refer to the preferred center, contrary to the unmarked case.

Hegarty, Gundel, and Borthen (2001) assume that clausal propositions, facts, or situations are more accessible to reference with *it* if they have already been mentally represented by the addressee. If a speaker refers to that entity, it causes the addressee to reprocess that entity, which renders it more salient. Two examples from Hegarty, Gundel, and Borthen illustrate this. Example (28a) shows that *it* cannot immediately be used to refer to the situation. Instead, *it* can only refer to the snake here. In contrast, in Example (28b), reference to the situation by *that* is possible, and due to that prior mention, subsequent reference by *it* is also possible.

(28) **There was a snake on my desk.**

a. # **It** scared me.

b. **That** scared me. **It** scared my office-mate too.

The second example illustrates that *it* vs. *that* can be used to indicate prior beliefs. In Example (29), the alternative replies (29a) and (29b) imply different background knowledge. After the statement by speaker A, the fact that linguists earn less than psychologists is at least activated. In Example (29a), speaker B uses *that*, thereby signalling the activated cognitive status of the abstract entity. However, in Example (29b), speaker B′ uses *it*, thereby implying that she already knew about this fact (such that it must already have been activated before the statement by speaker A), and due to being mentioned by speaker A, it has become salient and an entity in focus.

(29) A: I just read that **linguists earn less than psychologists**.

a. B: **That**'s terrible!

b. B′: **It**'s terrible!

Finally, Poesio and Modjeska (2005) analyzed 112 instances of *this* and *these* (used as pronouns and determiners). Forty-nine percent referred to nominal antecedents and 17% to non-nominal ones. They observed that 75–80% of the instances referred to entities other than highest-ranked entity (CP).

Referring to the topic of the utterance. The pronoun *it* tends to refer to the topic of the conversation, whereas demonstratives tend to refer to more peripheral antecedents. Poesio and Modjeska (2005) in their study of 112 instances of *this* and *these* found evidence for the hypothesis that these anaphors are used to refer to entities other than the CB of the current utterance (this was supported by 61–65% of the instances) or the CB of the previous utterance (supported by 90–93%).[30]

3.2.4 Distance Between Anaphor and Antecedent. In anaphora resolution, an important factor that affects the accessibility of antecedents is the distance between the anaphor and antecedent. Recency plays an important role in anaphora resolution systems (Mitkov 2002; Poesio, Ponzetto, and Versley 2010). A short distance between the anaphor and the antecedent implies a smaller search space and a smaller number of competing antecedent candidates, whereas a long distance implies a larger search space with many competing antecedent candidates. The distance can be measured in terms of tokens, sentences, spatiotemporal proximity, or the number of edges between nodes in some discourse structure, such as those posited by Rhetorical Structure Theory (RST) (Mann and Thompson 1988).

The distance preferences vary according to the anaphoric expressions used. We now list some tentative suggestions about these preferences from the literature.

Linear distance: demonstrative pronouns vs. personal pronouns. Comparing pronominal instances of NA and non-NA anaphora, Byron (2003, pages 34–35) observed that the more semantic information a pronoun has, the larger the average distance to its antecedent. The average distance of all pronouns in one of her data sets was 8.81 words. The largest

30 Poesio and Modjeska (2005, Section 3.1) call the CB (discourse) focus as well as topic.

average distance occurred with gender-marked personal pronouns (*she, hers, her*: 9.84; *he, his, him*: 9.2), followed by neuter personal pronouns (*it, its, they, them, their*: 8.72). Demonstrative pronouns occur closest to their antecedents (*this, that, these, those*: 8.18). The differences between the types do not seem large, though. Instances of non-NA anaphora would fall into one of the last two classes.

Linear distance: pronominal demonstratives vs. this *NPs.* The pronominal demonstratives *this* and *that* and the personal pronoun *it* are typically closer to their antecedents than *this* NPs (Schmid 2000; Kolhatkar 2015). In particular, demonstrative pronouns on their own are not particularly informative, and so the distance between the anaphor and the antecedent is fairly small and the textual coherence fairly strong (i.e., there are fewer competing candidates). In contrast, *this* NPs are informative because they are headed by a content noun. They license long-distance as well as short-distance antecedents, as shown in the following examples.

(30) Once an international poverty line is set, it must be converted to local currencies. This is trickier than it sounds. Currency exchange rates are inappropriate because most of the items that the poor consume are not traded on world markets. **Living expenses are much lower in rural India than in New York**, but **this fact** is not fully captured if prices are converted with currency exchange rates. (NYT)

Here, the distance between the anaphor and the antecedent is small: The antecedent of *this fact* occurs in the preceding clause. In contrast, in Example (31), the antecedent of *this question* occurs four sentences away from the anaphor sentence.[31]

(31) Among Roman Catholics, the differences were even more striking. Only 28 percent of Catholics who said religion was very or extremely important to them favored keeping abortion legal, but 72 percent of Catholics for whom religion was less important favored the legal status quo.

The sense of a public struggling with a morally difficult issue was dramatically conveyed when the survey asked: "**Would you approve or disapprove of someone you know having an abortion?**"

Thirty-nine percent said they would approve and 32 percent said they would disapprove. But 25 percent more volunteered a response not included in the question: They said their view would depend on the circumstances involved. An additional 5 percent did not know. The lack of a clear majority for either of the unequivocal responses to **this question** may be the best indicator of where public opinion really stands on abortion. (NYT)

Temporal and spatial distance: this *vs.* that. Among demonstrative pronouns, *this* tends to be associated with the entities that are spatially, temporally, or textually close to the speaker, whereas *that* tends to be associated with the entities that are not close to the speaker (Halliday and Hasan 1976). Textual proximity can be defined in terms of the relation between participants in a dialogue, that is, something said by the speaker versus something said by an interlocutor (Halliday and Hasan 1976). For instance, imagine a dialogue between two speakers, as shown in Example (32). Note how speaker A uses

31 In this example, one might argue that the question is implicitly stated in each intervening sentence. But from a computational perspective the explicitly stated clear antecedent of *this question* is four sentences away.

the pronoun *this* to refer to their own statement (i.e., *it's time we take action*), whereas speaker B uses *that* to refer to the same statement.

(32) A. **It's time we take action**. I know I have said **this** before, but now I really mean it.
B. What do you mean by **that**?

Example (33), from Halliday and Hasan (1976), demonstrates temporal proximity preferences: *That* tends to be associated with a past-time referent, whereas *this* for one in the present or immediate future.

(33) a. **We went to the opera last night**. **That** was our first outing for months.
b. **We're going to the opera tonight**. **This**'ll be our first outing for months.

Distance in a discourse structure. There are suggestions in the literature regarding the accessibility of antecedents in a discourse representation. This constraint is typically referred to as the **right-frontier constraint** (Polanyi 1985; Webber 1991) or the principle of availability (Asher 1993, page 313). The formal definition of the constraint varies according to the discourse representation theory and structure under consideration. That said, the general idea is that only those discourse segments can yield referents for anaphors that correspond to nodes on the right frontier of a formal discourse tree (Polanyi 1985; Webber 1991; Asher 1993, page 270; Asher 2008). The constraint is a visual representation of which salient nodes in a given discourse are accessible for later reference. The intuition is that given a discourse structure, represented as a tree, a referring expression cannot attach to a constituent to the left of the current constituent. An example from Webber (1991) is given in Example (34).

(34) a. There's two houses you might be interested in.
b. House A is in Palo Alto. It's got 3 bedrooms and 2 baths, and was built in 1950. It's on a quarter acre, with a lovely garden, and the owner is asking $425K. But **that**'s all I know about it.
c. House B is in Portola Valley. It's got 3 bedrooms, 4 baths and a kidney-shaped pool, and was also built in 1950. It's on 4 acres of steep wooded slope, with a view of the mountains. The owner is asking $600K. I heard all **this** from a real-estate friend of mine.
d. Is **that** enough information for you to decide which to look at?
e. # But **that**'s all I know about House A.

Here, parts (b) and (c) are central parts of the text. According to Webber, the continuation (e) is ill-formed, because, at this point, the information about House A is closed off and no longer accessible. The only accessible antecedents are the ones on the right frontier: (1) the information on both houses, that is, the information spanned by the root node and (2) the information on House B. Figure 3, from Webber (1991), illustrates the structure of the discourse tree at that stage. Only the nodes on the right side of the tree can serve as attachment points for (e).

Asher (1993, 2008) and Afantenos and Asher (2010) present a version of right-frontier constraint in the Segmented Discourse Representation Theory (SDRT) framework. They even demonstrate that SDRT's version of this constraint is respected about 95% of the time in their corpus of texts in French from different genres.

From a computational linguistics perspective, there are two problems with this constraint. First, researchers have demonstrated violations of the constraint (e.g., Poesio,

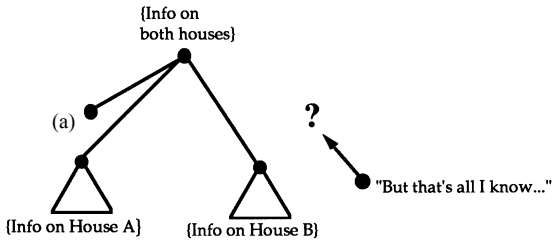


Figure 3
An illustration of the right-frontier constraint adapted from Webber (1991).

Patel, and Di Eugenio 2006). For instance, replacing either (d) or (e) in Example (34) with the (constructed) sentence in (e′) will violate the right-frontier constraint, because *that* in (e′) accesses the closed-off information about House A. However, it seems to be a fairly natural continuation of the conversation, especially if it were uttered in the course of a spontaneous conversation that had not been prepared in detail in advance.

e′. **That**'s all I know about House A, but I can give you more information about House B if you are interested.

Leaving aside this general question about the strictness of the right-frontier constraint, there is a more significant concern from an implementation perspective. It relates to the fact that non-NAs can be of various syntactic shapes. This means that the discourse representations derived from the state-of-the-art discourse parsers in computational linguistics such as those of Joty et al. (2013) and Feng and Hirst (2014), which are based on clauses as their elementary discourse units, will not always correspond smoothly with the syntactic shape or size of all non-NAs. That is, non-NAs may include other syntactic shapes such as verb phrases that would not be accessible in such discourse representations.

3.2.5 Underspecification of Non-nominal Antecedents. Non-nominal antecedents are typically not clearly delimited and identifiable stretches of discourse. Recasens (2008) discusses the non-specific nature of non-NAs, extending Poesio et al.'s (2006) Justified Sloppiness Hypothesis to non-NA anaphora. Poesio et al. argue with regard to nominal anaphora that the interpretation of some pronouns is not fully specified but is only "good enough" for the listener's purposes. For instance, in Example (35), it is not clear whether the antecedent of the pronoun *that* is the orange juice that has been loaded into the tanker car or the tanker car itself, and whether it matters.

(35) so then we'll
. . . we'll be in a position to
load the orange juice into the tanker car
. . . and send **that** off (Poesio et al. 2006, page 162)

Similarly, in Example (36), it is not clear whether *that* refers to postponing the vote, scheduling public hearings, rescheduling the vote, or to all of the above, and probably it is not necessary to understand the exact interpretation of *that* as far as the listener is concerned.

(36) Faced with complaints that the vote on adding bike lanes was being rushed, the municipal council opted to postpone the vote, to schedule public hearings, and to reschedule the vote for a later date, after every interested party had a chance to voice their concerns. The council hoped **that** would placate those who had expressed the most anger.

Poesio et al. (2006) propose four possible interpretations of such anaphors in the context of nominal anaphora: the two subparts of the antecedent, the composite object made from the two subparts, or an underspecified version of the full antecedent. Recasens (2008) suggests three possible interpretations for non-NAs: the largest or maximal discourse segment, the shortest or minimal discourse segment, and any discourse segment occurring between the largest and smallest discourse segments. We will see later in Section 4.2.6 that some approaches to non-NA anaphora keep the antecedent underspecified and mark just the verbal head as a proxy for it.

3.3 Summary and Outlook

In this section we discussed how non-NA anaphora is typically realized linguistically. Non-NA anaphora is usually realized with the pronoun *it*, demonstrative pronouns, and shell noun phrases (e.g., *this issue*). Not all instances of these expressions constitute instances of non-NA anaphors, and it is not straightforward for computational systems to identify whether a given instance is in fact an instance of a non-NA anaphor or not.

We also discussed some important properties of the phenomenon that have been addressed in the linguistics literature. A number of lexical, semantic, syntactic, and discourse-level preferences and constraints are associated with non-NA anaphora: The forms of the anaphor and the antecedent impose syntactic and semantic preferences on the antecedents. In general, demonstrative pronouns are preferred over the pronoun *it* when referring to non-NAs (Halliday and Hasan 1976; Webber 1979; Passonneau 1989). The pronoun *it* is typically used to refer to events, whereas the demonstratives *this* and *that* typically refer to propositions, facts, or situations (Hegarty, Gundel, and Borthen 2001; Gundel, Hedberg, and Zacharski 2005). The context of the anaphor also imposes preferences on the antecedents, and the context of the non-NAs imposes restrictions on the form of the anaphors that can be used to refer to them. For instance, at the lexical level, certain lexical items (e.g., *X is **true***) are suggestive of the syntactic and semantic type of the antecedent. Non-NA anaphora is primarily a discourse phenomenon, and some intuitions about the accessibility of different types of antecedents are described by Gundel, Hegarty, and Borthen (2003). The preferences for the distance between anaphor and antecedent vary according to the anaphoric expression. For instance, the demonstratives *this* and *that* differ with respect to the relative location of the entity they refer to. *This* tends to refer to an object near the speaker, whereas *that* is preferred for more distant referents. Distance, in this case, could refer to temporal distance, textual location, or the relation between participants in a dialogue, namely, something said by the speaker versus something said by an interlocutor (Halliday and Hasan 1976). In general, non-NAs are underspecified, and it is not always clear what the ground truth antecedent is for a given anaphor instance, and this poses a serious challenge to computational systems.

Many of the insights presented in this section stem from observations based on corpora that have been annotated with the respective features. In the next section, we will discuss such annotation efforts carried out for non-NA anaphora.

4. Annotation of Non-NA Anaphora

Anaphoric expressions with non-NAs have been a subject of interest for the last few decades, and corpus-based studies for English have been carried out in a variety of domains and registers, from news texts (Botley 2006; Kolhatkar, Zinsmeister, and Hirst 2013a; Uryupina et al. 2018) to academic articles (Kolhatkar and Hirst 2012), to narratives (Botley 2006; Uryupina et al. 2018), to spoken monologues (Botley 2006; Guillou et al. 2014), to spoken dialogues (Schiffman 1985; Eckert and Strube 2000; Byron 2003; Uryupina et al. 2018) or multi-party discussions (Müller 2008). We first highlight some of the challenges that annotation efforts have faced and then present selected important resources in more detail.

4.1 Challenges Associated with Annotating Non-NA Anaphora

Annotating non-NA anaphora is a challenging problem and researchers have discussed associated difficulties, low inter-annotator agreement, and how they worked around the difficulties.

Classifying an anaphor as an instance of non-NA anaphora. Halliday and Hasan (1976, pages 66–68) informally analyzed 51 demonstratives (determiners and pronouns) in the last two chapters of *Alice's Adventures in Wonderland* as part of their investigations of text cohesion. They mention that it is not always easy to distinguish between NA and non-NA uses.

Identifying suitable non-nominal antecedents. Botley and McEnery (2001) and Botley (2006) make the point that non-NA anaphora poses difficulties for corpus-based linguistics in that 29% of the cases (186 instances) were hard to analyze. Botley (2006) points out two main reasons for this difficulty: the lack of clear surface linguistic boundaries and the complex or unclear inference process for retrieving antecedents.

Poesio and Modjeska (2002, 2005) analyzed 112 *this* NPs. They were interested in the cognitive status of *this* NPs in the given discourse. Because of the difficulties associated with identifying the precise antecedents of *this* NPs, they developed an annotation scheme where the annotators do not have to mark the actual antecedents. Rather, the scheme instructs the annotators to classify *this* NPs into different categories such as visual deixis, discourse deixis, and anaphoric, and, based on these categories, they assign a cognitive status to each *this*-NPs instance. The annotators achieved agreement of $\kappa = 0.82$ in this classification task.[32]

Artstein and Poesio (2006) report two experiments where 20 untrained annotators were asked to mark antecedents of NA and non-NA anaphora in a TRAINS91 dialog (Allen and Heeman 1995).[33] In the first experiment, they asked naive annotators to mark unconstrained regions of text as antecedents for NPs in general, including non-NA anaphors. In the second experiment, four annotators with prior experience annotated another dialog. This time, only (sets of) entire utterances could be marked as the antecedent. In the first experiment, in only 42% of the cases did annotators agree with the most popular choice for the beginning of the antecedent, and in 64% they agreed with

32 Poesio and Modjeska (2002) use Fleiss' (1971) κ (M. Poesio, personal communication).
33 `http://www.cs.rochester.edu/research/cisd/resources/trains.html`.

the most popular choice for the end. The second experiment showed a similar tendency for annotators to agree more on the ends of the segments than on their beginnings.

Determining semantic and cognitive features. Gundel, Hedberg, and Zacharski (2004) analyzed 99 instances of demonstrative pronouns. Initially, they planned to annotate the semantic types of both the anaphor and its antecedent (which might differ because of coercion). They assumed that there are clear correlations between the syntactic form and the semantic type of an antecedent (e.g., VPs denote either activities or states). In contrast, the semantic type of the anaphor is hard to determine and can only in certain cases be deduced easily from the predicate's semantic restrictions. They therefore abandoned annotating exact semantic types and instead marked the relation as direct when the referent of the anaphor was the same as the referent of the antecedent and indirect otherwise. They classified the pronouns into six categories: nominal direct, nominal indirect, non-nominal direct, non-nominal indirect, pleonastic, and other. All three coders agreed on the classification of only 56 out of 99 instances.

Hedberg, Gundel, and Zacharski (2007) analyzed 321 instances of demonstrative pronouns and report $\kappa = 0.46$ (moderate agreement) for identifying the cognitive status of the antecedent (activated or in focus), and $\kappa = 0.70$ (substantial agreement) for identifying the type of the antecedent (direct or indirect).[34] They do not report agreement in identifying the actual antecedents.

4.2 Resources

Synthesizing annotation approaches related to non-NA anaphora is not trivial, as these approaches (a) do not always share the same goals, (b) consider different anaphoric expressions in different corpora and domains, or (c) focus on different properties. A survey of annotation efforts for non-NA anaphora in different languages has been written by Dipper and Zinsmeister (2012), and we do not want to repeat that information in this article. Instead, in this section, we will focus on corpora that provide useful information for building computational systems. In particular, we will focus on the following questions.

1. What is the goal of the annotation?
2. What is covered by the annotation? This question concerns three aspects: First, which anaphoric expressions are considered for annotation? Second, are antecedents marked and, if so, how does the annotation approach the challenge of annotating text segments with imprecise boundaries? Third, how are other NPs annotated, for example, non-referring NPs or NPs that do not co-refer with some other expression (*singletons*). The last aspect is relevant to all systems that deal with naturally occurring data, where the task is first to classify an NP as referring or not, and second, to classify a referring NP as an instance of a NA or non-NA anaphor.
3. What is the size of the resource (number of tokens, number of non-NA instances), and which text types or domains are considered?
4. Which other features are included in the annotation?

34 Hedberg, Gundel, and Zacharski (2007) use Siegel's (1993) K (R. Zacharski, personal communication), which is identical to Scott's (1955) π.

5. What documentation is available? How do they measure the inter-annotator agreement?
6. Which corpora are available for researchers and what information from the annotations can be incorporated into computational systems? To what extent can the corpora be used as training data for machine learning systems?

For coreference in general, there are quite a few annotated corpora available. Most of them, however, focus on what Uryupina et al. (2018) call "relatively easy cases" of anaphoric reference, that is, anaphora with nominal antecedents. As described in the previous section, annotating non-NA anaphora represents a major challenge and is often done by trained expert annotators. Hence, corpora annotated with non-NA relations are scarce and often small. In this section, the most important resources of non-NA anaphora in terms of size and richness of annotation are described in detail.[35]

4.2.1 Schiffman (1985); Passonneau (1989). Passonneau (née Schiffman) investigated the use of pronominal *it* and *that*. The assumption underlying her work is that the two pronouns serve complementary discourse functions: In certain contexts, the anaphor *it* will be the unmarked (i.e., most frequently chosen) option and *that* the marked one, and vice versa. Schiffman (1985) and Passonneau (1989) cover both NA and non-NA anaphora.

In her thesis, Schiffman assumed that conversational data (in particular, conversations involving problem-solving, information-gathering, or counseling situations) would contain a high frequency of anaphoric *it* and *that*. She decided to examine conversations involving speakers from similar backgrounds to increase the degree of uniformity across different individuals, and the participants should not know one another so that their conversations would be easy to follow for observers. This resulted in a corpus of career-counseling interviews between a professional counselor and graduate students at the university.

From eight interviews, four were randomly selected for the current study, annotated by a trained student research assistant, and checked several times by the author of the study. In total, the corpus contained 31,780 tokens, among them 983 instances of pronominal *it* and *that*, of which 298 were instances of non-NA anaphora.

In her study, Schiffman (1985) distinguished between three kinds of anaphoric relations, depending on the form of the antecedent: NPs, VPs, and sentential constituents (clauses, sentences, and paragraphs).[36] In her work, the term NP is defined very broadly and includes nominalizations in Vendler's (1968) sense, such as gerunds or *that* clauses. This makes it rather difficult to extract results from her thesis that concern just the instances of non-NA anaphora according to our definition. At some point in her analysis, however, she distinguishes between different types of NP antecedents, which can be ordered according to their "nouniness," from canonical, "true NPs" with a lexical head to derived nominals and gerunds to sentence-like NPs in the form of infinitives and clauses introduced by *that*, *whether*, or *if*. The categories that are the least noun-like (infinitives and clauses) would be instances of non-NA anaphora by our definition.

Schiffman (1985) considered a range of syntactic and pragmatic factors, with a focus on directly observable, theory-independent aspects. Syntactic features include

35 At `https://github.com/kvarada/non-NA_Resources` we maintain a repository that collects all kinds of resources related to non-NA anaphora, including (links to) corpora and annotation guidelines.

36 Schiffman uses the term *nominal anaphora* to refer to anaphora with NP antecedents, and *non-nominal anaphora* for anaphora with VP antecedents.

the clause level (if it occurs in a main or subordinate clause) of the anaphor and its antecedent, if present, and its grammatical function (subject, direct object, copula predicate,[37] other). Another feature concerned the form of the antecedent, if present, with the possible values NP, VP, S, none (mainly for non-referential uses), discontinuous, and multiple (for ambiguous cases). Further features described the relation between the anaphor and the antecedent: sentence distance (same, adjacent, remote sentence) and adjacency, recording whether some other NP intervenes between the anaphor and the antecedent that could serve as an antecedent. Possible values of adjacency were: adjacent (no intervening NP), nearest semantic match (no inanimate, singular NP intervenes), not adjacent (at least one inanimate, singular NP intervenes).

Schiffman (1985) compares the use of *it* and *that* with a statistical analysis of the annotated features. The results show certain preferences for one of the pronouns, depending on the antecedent's form: If the antecedent is non-nominal, *that* is preferred; Otherwise, if the antecedent is a pronoun, *it* is preferred. If the antecedent is a (true) NP, *it* and *that* are equally likely (Schiffman 1985, Section 5.5.2). In such cases, the grammatical function of the antecedent plays a role: If it is a subject, *it* is preferred; otherwise, *that* is more likely.

4.2.2 Eckert and Strube (2000). These researchers annotated data from spoken language with the goal of evaluating the performance of their anaphora resolution algorithm. They chose spoken rather than written data because they expected spoken data to contain more pronominal anaphors and a more diverse range of anaphor types, including non-NA anaphors and vague anaphors, which lack a clearly defined linguistic antecedent. The corpus they analyze is the Switchboard corpus (Godfrey and Holliman 1993),[38] which contains transcribed telephone conversations between two people who were not acquainted with each other and were asked to talk about given topics such as childcare. The authors argue that this sort of data is easier to follow than unconstrained conversations, and, at the same time, it is more diverse than task-oriented dialogues like TRAINS (Allen and Heeman 1995),[39] which often contain a lot of imperative-like constructions and mostly refer to concrete objects.

Eckert and Strube (2000) randomly selected five dialogues from the Switchboard corpus. Two were used for training the annotators and the other three for calculating inter-annotator agreement and evaluating their resolution algorithm.

The corpus consists of 8,421 dialogue act units, which roughly correspond to main clauses plus any subordinate clauses. The annotators analyzed personal pronouns (527 instances) and demonstrative pronouns (151 instances). They did not mark first and second-person pronouns, or non-referring pronouns, such as expletives (pleonastic *it*). Among the annotated pronouns, only 45.1% referred to nominal antecedents and 22.6% to non-NAs. In addition to these two classes, they define two further classes: vague anaphors and inferrable-evoked pronouns. Vague anaphors are pronouns without a clearly defined linguistic antecedent, which, for example, refer to the general discourse topic (13.2% of the pronouns). Inferrable-evoked pronouns are particular uses of plural *they* without explicit antecedents and whose referent has to be inferred; these are never instances of non-NA anaphors (19.1%). For the classification task (nominal,

37 Schiffman (1985) calls this function predicate. For anaphors, she provides as examples *That's about it/that, There's this/that*.

38 `https://catalog.ldc.upenn.edu/ldc97s62`.

39 `http://www.cs.rochester.edu/research/cisd/resources/trains.html`.

non-nominal, vague, inferrable-evoked), Eckert and Strube report an inter-annotator agreement of Fleiss' (1971) $\kappa = 0.81$ for personal pronouns, and $\kappa = 0.80$ for demonstrative pronouns.

The subset of anaphors that were classified in the same way by both annotators was further annotated in that both annotators marked the antecedents and then agreed upon a reconciled version of the data. Non-nominal antecedents consisted of only VPs and clausal antecedents. Annotator accuracy for antecedent marking was measured against the reconciled version. For non-NA anaphora, accuracy (percent agreement) of the two annotators was 85.7% (60 correct of 70 cases) and 94.3% (66 correct), respectively (as compared with NA anaphora with accuracies of 96.1% and 98.4%, respectively).

The annotation procedure is documented in Eckert and Strube (2000), but the annotated corpus is not available.

4.2.3 Botley and McEnery (2001); Botley (2006). The goal of Botley and McEnery (2001) and Botley (2006) was to investigate the use of demonstratives from a corpus linguistics perspective and, in particular, to investigate the different usages of such anaphora across different genres. They examined three corpus samples, each with 100,000 words: the Associated Press corpus (AP) of American newswire texts, the Hansard corpus with proceedings from the Canadian House of Commons, and the American Printing House for the Blind corpus (APHB), a collection of literary works and motivational narrative.[40] They expected that anaphors would function differently in the three corpora. For instance, the Hansard corpus contains spoken data, in particular, exchanges between parliamentarians who refer to each other's arguments as well as their own using non-NA anaphora. Whereas the Hansard sample is a continuous record of one parliamentary session, the AP and APHB samples contain a range of texts dealing with a variety of topics.

They considered both demonstrative pronouns and *this* NPs in their study. In addition to non-NA anaphora, they also investigated instances of pure textual deixis (see Sections 2.1.3 and 2.1.5 for more on their notion of indirect anaphora).

The three corpus samples contained 403 non-NA demonstratives, of which 57% occurred in the Hansard sample, 29% in the APHB sample, and 14% in the AP sample, providing evidence that spoken data exhibits more instances of non-NA anaphora than written data.

The corpus was annotated by the authors without any reliability measurement, and the data itself is no longer available (S. Botley, personal communication). The annotations included the type of anaphora (non-NA anaphora, NA anaphora, exophoric), direction of reference (anaphoric, cataphoric), the type of the anaphor (*this* NP, pronoun, not applicable),[41] and the type of the antecedent (nominal, clausal, propositional/factual,[42] adjectival, none). Instances of non-NA anaphora fall into the classes of clausal and propositional antecedents.

Botley and McEnery (2001) perform a series of statistical tests, combining the individual forms (*this, that, these, those*) with each of the annotated features. Figures on non-NA anaphora cannot easily be extracted from most of the tables, though,

40 `http://ucrel.lancs.ac.uk/corpora.html`.

41 In Botley and McEnery's (2001) terms: noun modifier (for determiner, within a shell noun phrase) and noun head (for pronoun).

42 These are "surface statements or utterances made by speakers or writers" (Botley 2006, page 11), i.e., probably direct speech.

because they include instances of nominal indirect anaphora, also called bridging (see footnote 6).

4.2.4 Byron (2003). Byron's (2003) primary goal was to compare and contrast the use of personal vs. demonstrative pronouns, and her secondary goal was to compare and contrast their use in two different genres. These investigations were intended to aid in the development of an automated system that could deal with and interpret anaphora in natural language, including non-NA anaphora. She performed two extensive annotation studies, both based on spoken data, which contain a larger number of non-NA instances than written data. The first study was based on the TRAINS93 corpus (Allen and Heeman 1995),[43] which consists of task-oriented spoken dialogues. The dialogues involve two participants, one person who has a certain task to accomplish, involving the routing and scheduling of freight trains, and one person who helps with the planning. The second study used the Boston University Radio Speech corpus (Ostendorf, Price, and Shattuck-Hufnagel 1996)[44] (BUR), comprising transcripts of news stories read over the radio.

For the second study, the annotation scheme was modified in various places, wherever the annotators had difficulties in the first annotation task; the annotators were two students without prior knowledge of the topic and the author of the study. For instance, the initial guidelines used the term *linguistic antecedent*, which turned out difficult to work with. The antecedent was then called *linguistic trigger* and finally *linguistic anchor*. The new genre (prepared spoken monologues rather than spontaneous dialogues) also required some adaptation of the guidelines. For instance, the scheme of TRAINS93 only provided two values for grammatical function (subject and non-subject). The BUR corpus contained more elaborate structures than TRAINS93, so the grammatical function attribute was subdivided into more fine-grained values (subject, direct object, indirect object, object of a preposition, possessive, and other). Similarly, though the initial scheme distinguished between four different values for the syntactic category of the antecedent (NP, pronoun, non-NP, and none), the final scheme had eight values (NP, pronoun, non-NP, none, name, title, noun modifier, and possessive).

Byron focused on referring singular and plural neuter personal and demonstrative pronouns (*it, its, itself, them, themselves, they, their, this, that, these,* and *those*). The BUR annotation additionally included *he, his, him, himself, she, her,* and *herself* (none of these forms occurred in TRAINS93). For TRAINS93, the annotators were trained on one dialogue, with subsequent corrections of the annotation guidelines. They then analyzed pre-marked referring pronouns in 19 randomly selected dialogues, with a total of 10,420 tokens, among them 347 relevant pronouns. For BUR, they used two monologues for training, and analyzed 35 monologues. The monologues were selected such that each monologue contained at least one demonstrative pronoun, hence, the BUR figures are not representative of the corpus. BUR has a total of 13,415 tokens, among them 380 relevant pronouns.

Byron annotated 190 instances of *it* or *its* (122 from TRAINS93 corpus and 68 from BUR corpus), and 227 instances of demonstrative pronouns (177 from TRAINS93 and 50 from BUR corpus). In TRAINS93, 11 (7%) of the personal and 29 (32%) of the demonstrative pronouns are instances of non-NA anaphors; in BUR, 5 (2%) of the personal and 23 (46%) of the demonstrative pronouns.

43 http://www.cs.rochester.edu/research/cisd/resources/trains.html.
44 https://catalog.ldc.upenn.edu/LDC96S36.

Byron's (2003) annotation scheme is based on Schiffman's (1985) annotation scheme (see Section 4.2.1). In particular, she used a variety of syntactic features of Schiffman's scheme: the clause level of the anaphor and the antecedent, their grammatical function, the distance between the anaphor and its antecedent, and the antecedent's syntactic category. In TRAINS93, possible antecedents had to be contained within the same utterance, to prevent the annotators from marking huge blocks of discourse as the antecedent, for example, if reference was made to the plan under discussion. In BUR, in contrast, marking antecedents in a previous paragraph was allowed.

In addition to Schiffman's features, Byron's (2003) scheme also aims at specifying properties of the anaphor's referent, which is called its *semantic antecedent*. According to Byron (2003, page 12) "the semantic antecedent is your gut feeling about what the pronoun is standing in for." She suggests, as a test, substituting (an appropriate expression realizing) the semantic antecedent for the pronoun to see if the meaning remains constant. An example from Byron (2003, page 13) involving non-NA anaphora is provided in Example (37). The anaphor and the (linguistic) antecedent are marked as usual, the semantic antecedent of *that* is the *fact* or the *proposition* or the *idea* that an engine cannot arrive at Bath in time. The (linguistic) antecedent is then a previous mention of the semantic antecedent, possibly using different words.

(37) u: . . . **we can't get an engine to Bath in time**
s: **That**'s right

Byron reports inter-annotator agreement of two annotators in terms of κ value.[45] For annotating the semantic antecedent, the κ value was 0.82 (TRAINS) and 0.86 (BUR) for personal pronouns, and 0.56 (TRAINS93) and 0.53 (BUR) for demonstrative pronouns. For the (linguistic) antecedent, only those cases were considered where both annotators agreed on the semantic antecedent. The κ value was 0.77 (TRAINS93) and 0.95 (BUR) for personal pronouns, and 0.37 (TRAINS93) and 0.62 (BUR) for demonstratives. Personal pronouns were easier to annotate than demonstratives with respect to these fields because the majority of these instances are usually NA anaphora. (Linguistic) antecedents with demonstratives showed very low agreement in TRAINS93. This is partly because one category dominates ('no antecedent' in 49%), and expected agreement is high, resulting in a low κ value.

The corpora are freely available.[46] Detailed documentation of the annotation is provided in Byron (2003).

4.2.5 Pradhan et al. (2007): OntoNotes. Starting in 2006, the OntoNotes project, with members from BBN Technologies, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences Institute, has worked over a period of several years to create a large, multilingual resource called OntoNotes. The corpus comprises various genres (telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs) in three languages (English, Chinese, and Arabic). The final release, OntoNotes Release 5.0, was published in 2013 and contains about 1.5 million tokens. The corpus builds on other resources, including the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993)[47] and the Proposition Bank (Palmer, Kingsbury, and Gildea 2005),[48] and is richly

45 Byron (2003) uses Scott's π (Scott 1955).
46 The data are available from `https://github.com/kvarada/non-NA_Resources`.
47 `https://catalog.ldc.upenn.edu/LDC99T42`.
48 `https://propbank.github.io/`.

annotated with syntax and predicate-argument structure, word senses, and coreference. Because to its size and rich annotations, OntoNotes is very popular and has been used in several CoNLL shared tasks.

The focus of OntoNotes is on NA anaphora. In addition, events are also considered, if they are mentioned again either by a pronominal anaphor or instances of certain event-denoting NPs (e.g., nominals derived from verbs such as *growth* or *contribution*). In OntoNotes, only the verbal head of the event-denoting expression that serves as the antecedent is annotated; see Example (38), from Pradhan et al. (2007).

(38) Sales of passenger cars **grew** 22%. **The strong growth** followed year-to-year increases.

According to Chen, Su, and Tan (2010), Release 2.0 of OntoNotes contained 1,235 anaphora referring to an event, which were 19.97% of all anaphora annotated in the corpus. Among them, 59.35% were instances of event-NP anaphors (like *growth*), and 40.65% were instances of the pronominal anaphors *this* (16.9%), *that* (54.1%), and *it* (29.0%).

The OntoNote creators state that they strove to ensure that each annotation layer has a least 90% inter-annotator agreement (Weischedel et al. 2013, page 4). OntoNotes is available via LDC, along with documents describing the annotation guidelines.[49]

A similar enterprise is the task of event coreference annotation, which requires identification of co-referring event verbs, as shown in Example (39), taken from Lee et al. (2012).

(39) The New Orleans Saints **placed** Reggie Bush on the injured list on Wednesday. Saints **put** Bush on I.R.

Cases of reference to events as shown in Examples (38) and (39) clearly differ from non-NA anaphora as considered in this survey, in that the anaphor is either a verb or the anaphoric noun does not allow for the usual lexico-grammatical patterns of shell nouns (see Section 3.1.2).

4.2.6 Müller (2008). Müller deals with the automatic resolution of the pronouns *it*, *this*, and *that* in unrestricted multi-party dialogue as part of a summarization system producing extracts as summaries, where unresolved pronouns often constitute a problem. He aims to create a system that is usable in a real-world setting, so no manual preprocessing (such as filtering non-referring pronouns) is involved. However, the system's input are manually created transcripts with correct spelling.

Müller (2008) used the ICSI Meeting Corpus (Dhillon et al. 2004),[50] which was produced as part of a project called Meeting Recorder. It is a collection of transcripts of group discussions during project meetings involving 3 to 10 speakers, dealing with rather technical topics. For the study, Müller randomly selected five dialogues dealing with different topics (natural language understanding, the Meeting Recorder project itself, the Internet and networking, signal processing, and robustness for speech recognition). The study had three goals: (a) Examining the distribution of different usages of *this*, *that*, and *it* in this multi-party spoken dialogue domain; (b) automatically detecting non-referential *it*; and (c) automatically identifying antecedents of referential instances of *this*, *that*, and *it*. With regard to non-NA anaphora, Müller (2008) made the simplifying assumption that only VPs can be non-NAs.

49 https://catalog.ldc.upenn.edu/ldc2013t19.
50 http://www1.icsi.berkeley.edu/Speech/mr/.

It is unclear how many tokens the annotated corpus consists of. It contains at most 150 instances of a subset of non-NA anaphora (depending on the setting; see the following description).

Müller carried out two annotation experiments. The first focused on classifying instances of *it*, *this*, and *that* into six classes. The main class (normal) contains NA and non-NA anaphors. The other classes cover vague pronouns (those without an identifiable antecedent, referring, e.g., to the general topic), discarded pronouns (if an utterance is abandoned), the pronoun *it* in specific syntactic constructions (two classes: with extraposition and as a syntactic filler), and one class for other uses. The first two classes, normal and vague, are considered subtypes of referential *it*, and the other four are considered subtypes of non-referential *it*.

The first task was performed by two naive annotators, non-native speakers of English, who annotated all five dialogues.[51] Müller (2008) only reports agreement results for 1,040 instances of the pronoun *it*. The annotators achieve a value of $\kappa = 0.64$ for the main class (normal), containing NA and non-NA *it*.[52] After the annotation, the annotators created a reconciled gold version.

The second experiment focused on identifying antecedents of referential instances of *this*, *that*, and *it* from the first experiment, in the form of pronouns, (full) NPs, or VPs. Müller chose to work with naive annotators, which precluded the use of sophisticated annotation schemes. For NP antecedents, the annotators were instructed to mark the head of the NP plus any premodifiers, but excluding any postmodification. For VP antecedents, they were instructed to mark only the head of the VP. He chose this approach to improve inter-annotator agreement and to make it easier to develop an automated system for this task. The second task involved the two annotators from the first experiment plus two new annotators, who were native speakers of English.

Evaluating agreement concerned two aspects: identifying some text span as an antecedent and linking anaphors to their antecedents. On average, all four annotators agreed on some text span as an antecedent only 27.77% of the time. Antecedents could be the pronouns *it*, *this*, or *that* (occurring in a chain of anaphors), NPs, or verbal heads. In particular, Müller observed that annotators encountered the most difficulty in identifying NP and verbal-head antecedents: On average, all four annotators agreed on such antecedents only 6.06% of the time. With regard to anaphoric linking, Müller reports chance-corrected agreement in terms of a variant of Krippendorff's (2004) α described in Passonneau (2004). This metric requires all annotations to contain the same set of antecedents. So he computed Krippendorff's α on the intersection of antecedents found by all annotators—that is, a rather small set in the case of non-pronominal antecedents. For linking these non-pronominal antecedents, the inter-annotator agreement was in the range of $\alpha = 0.70$–0.88.

Instead of having the annotators create a reconciled version, Müller (2008) automatically generates a core data set, based on the four annotations. He creates three different sets, one with all anaphoric links that at least two annotators agreed on ($n = 2$), one with all links from three annotators ($n = 3$), and one with the links from all four annotators ($n = 4$). In all settings, VP antecedents are infrequent: 16.84% of the links (150 instances) with $n = 2$, 12.24% for $n = 3$, and 6.38% for $n = 4$. The drop in the proportion with

51 Müller analyzed the initial annotations and corrected problems in the annotation scheme. The same annotators annotated the same dialogues again, according to the modified scheme. Inter-annotator agreement refers to the second annotation.

52 Müller (2008) uses Fleiss' measure (Fleiss 1971), which amounts to Scott's π in the case of two annotators.

increasing n can be taken to indicate that these cases are more difficult to annotate than other instances of anaphora.

The annotated corpus is not available. Müller (2008) contains detailed documentation of the annotation.

4.2.7 Kolhatkar and Hirst (2012): This-issue corpus. Kolhatkar and Hirst annotate instances of non-NA anaphora realized by the shell noun phrase *this issue*. Their goal was to build training and test data that can be used by a machine learning system to resolve instances of *this issue*. They chose to work with *this issue* instances from MEDLINE abstracts[53] because (a) the antecedents of *this issue* are relatively well defined in this domain, (b) the limited context of abstracts restricts the antecedent search space, and (c) *issues* in MEDLINE abstracts are generally associated with clinical problems in the medical domain, and the extraction of this information would therefore be useful in any biomedical information retrieval system.

Their corpus contains 183 instances of *this* modifier* *issue* (i.e., *this* followed by optional adjectives plus the noun *issue*) along with the surrounding context from MEDLINE abstracts. Of these instances, 132 instances were independently annotated by two annotators, a domain expert and a non-expert, and the remaining 51 instances were annotated only by the domain expert. The annotation task was to identify and mark automatically parsed constituents as antecedents, without concern for their syntactic types. The majority of antecedents were non-nominal: clauses (37.9%) or sentences (26.5%) or a sequence of adjacent constituents (18.2%).

Example (40) shows an annotated example from their corpus. The antecedent *that avoidance of nitrous oxide ...* is marked for the anaphor with ID="2". The REFERENT_TYPE of this antecedent is "CLAUSE" and the DIST attribute has the value "ADJA" as it lies in the adjacent sentence. The annotator included an EXTRA attribute of type PARAPHRASE in the annotation because the actual referent (which would be *whether avoidance of nitrous oxide ...*) is not explicitly stated in the text.

(40) From this preliminary study with a low statistical power, it appears <ANTECEDENT ID="2">**that avoidance of nitrous oxide in one's practice may not affect the outcome in the neurosurgical patients**</ANTECEDENT>. Further large systemic trials are needed to address <ANAPHOR ID="2" DET="this" NOUN="issue" REFERENT_TYPE="CLAUSE" DIST="ADJA" EXTRA="PARAPHRASE:whether avoidance of nitrous oxide in one's practice affects the outcome in the neurosurgical patients">**this issue**</ANAPHOR>.

Because the boundaries of such antecedents are fuzzy, Kolhatkar and Hirst argue that such annotations need an inter-annotator agreement coefficient that goes beyond the match/mismatch binary and incorporates distance between strings more elegantly than Krippendorff's α with conventional distance metrics. They use Krippendorff's unitizing α (α_u; Krippendorff 1995, 2004, 2013) and report an inter-annotator agreement of $\alpha_u = 0.86$, which is considered to be a strong indicator for reliably annotated data. The corpus and the annotation guidelines are available for non-commercial use.[54]

4.2.8 Kolhatkar and colleagues: The ASN and CSN corpora. Kolhatkar, Zinsmeister, and Hirst (2013a, 2013b), Kolhatkar and Hirst (2014), and Kolhatkar (2015) annotated two relevant

53 MEDLINE is a database of references and abstracts on life sciences and biomedical topics. https://www.nlm.nih.gov/bsd/pmresources.html.

54 https://github.com/kvarada/non-NA_Resources/tree/master/Kolhatkar-Hirst_2012.

corpora: the Anaphoric Shell Nouns (ASN) corpus and the Cataphora-like Shell Nouns (CSN) corpus.

The ASN corpus. Kolhatkar, Zinsmeister, and Hirst extended the annotation of *this issue* to other shell nouns in news domain. They created the ASN corpus, which comprises 1,810 anaphoric instances of six shell nouns: *fact*, *reason*, *issue*, *decision*, *question*, and *possibility* from the New York Times corpus (Sandhaus 2008),[55] and their corresponding antecedents. Similar to Kolhatkar and Hirst (2012), they chose shell noun instances following the pattern *this* modifier* shell noun (*this* followed by optional adjectives followed by a shell noun). With this corpus, the authors hoped (a) to learn to what extent non-experts can identify non-NAs, and (b) to create a corpus for the evaluation of their computational system that resolves anaphoric shell noun phrases.

For (a), they chose CrowdFlower[56] as their crowdsourcing platform and annotated 2,323 anaphoric shell noun instances. They divide the annotation task of marking non-NAs into two relatively simple sequential annotation tasks: identifying the sentence containing the antecedent and, given this sentence, identifying the antecedent segment in that sentence.[57] Each instance was annotated by eight annotators. Their final curated corpus contains 1,810 instances of anaphoric shell noun phrases and their antecedents.

They report a Krippendorff's unitizing α of 0.54 and a Krippendorff's α of 0.51 and 0.61, using the Jaccard and Dice distance metrics, respectively, indicating only moderate inter-annotator agreement. That said, considering that the crowdsourcing platforms are not designed for comprehensive annotation guidelines and the annotators are non-experts in the linguistic phenomenon, the results are encouraging, as they suggest that people do have more or less similar intuitions about non-NAs.

They observed two primary challenges in their annotation experiments. First, the question of "what to annotate" as mentioned by Fort, Nazarenko, and Rosset (2012) was not straightforward for ASN antecedents, as identifying the boundaries of the antecedents is more complicated in that case than with ordinary NA anaphora. And second, the notion of a *right* answer was not well defined for non-NAs, because the boundaries of antecedents are not always clearly delimited.

They also assess the quality of the crowd annotation on a sample of 300 instances with the help of experts. They asked two judges to rate the acceptability of the crowd-sourced answers based on the extent to which they provided the correct interpretation of the corresponding anaphor. They observed that 84.6% of the total instances were acceptable according to both judges.

The CSN corpus. The CSN corpus consists of about 114,700 cataphora-like instances of the same six shell nouns from the New York Times corpus (*fact*, *reason*, *issue*, *decision*, *question*, and *possibility*) and their shell content. Kolhatkar, Zinsmeister, and Hirst (2013b) aimed to use supervised machine learning methods to resolve anaphoric shell noun phrases. In order to do that, they started by automatically creating training data using the cataphora-like shell noun constructions from Table 3 in Section 3.1.2 and extracting the shell content using syntactic information. In particular, they extracted instances following seven cataphora-like patterns: *N-be-to*, *N-be-that*, *N-be-wh*, *N-to*,

55 `https://catalog.ldc.upenn.edu/ldc2008t19`.

56 Now at `https://www.figure-eight.com`.

57 Because the instances do not contain plural forms of demonstratives, it is unlikely that the antecedent spans more than one sentence.

N-that, *N-wh*, and *N-of*, and one anaphora-like pattern, *Sub-be-N*. Later, Kolhatkar and Hirst (2014) observed that simply using lexico-syntactic patterns creates noisy data. So they used Schmid's (2000) semantic frames to extract the semantic preferences of different shell nouns and incorporated this information when creating automatically labeled training data. To evaluate the quality of their automatically labeled data, they annotated about 100 instances of each of 12 shell nouns (*idea*, *issue*, *concept*, *decision*, *plan*, *policy*, *problem*, *trouble*, *difficulty*, *reason*, *fact*, and *phenomenon*) using crowdsourcing. They observed that three out of five annotators agreed on 1,152 instances.[58]

4.2.9 Uryupina et al. (2018): The ARRAU corpus. Uryupina et al. present the second release of the ARRAU corpus. ARRAU's goal is to serve as a corpus for "the next generation of coreference/anaphora resolution systems." The corpus encodes a large variety of linguistic information, so that advanced systems that combine such information with world knowledge can profit from the data.

In the ARRAU corpus, all NPs are annotated, including non-referring NPs and non-NA anaphors.[59] For both NA and non-NA anaphora, antecedents are also marked. ARRAU builds on other corpora, so that existing annotations can be reused. In particular, ARRAU integrates:

- A subcorpus of news texts called RST (around 230,000 tokens), which consists of the subset of the Penn Treebank that was annotated for the RST Discourse Treebank[60] (Carlson, Marcu, and Okurowski 2001). This subcorpus is already annotated with syntax (through the Penn Treebank), rhetorical structure (through the RST Discourse Treebank), and argument structure (through the PropBank).
- A subcorpus of spoken dialogues called TRAINS (85,000 tokens), which includes task-oriented dialogues from the TRAINS91 and TRAINS93 corpora.
- A subcorpus called GNOME (20,000 tokens), which consists of a subset of the GNOME corpus (Poesio 2000)[61] and includes descriptions of museum objects and brief texts about the artists that produced them, and a selection of pharmaceutical leaflets providing patients with mandatory information about their medications. The corpus is already annotated with discourse units and NA anaphors and their antecedents.
- A subcorpus of spoken narratives called PEAR (15,000 tokens), which includes all narratives in English from the Pear Stories project[62] (Chafe 1980). These are narratives by subjects who watched a film involving pears and then recounted its contents.

58 More details on these corpora can be found in Kolhatkar (2015, pages 86, 105) and Kolhatkar, Zinsmeister, and Hirst (2013b). Annotation guidelines are available at `https://github.com/kvarada/non-NA_Resources`. The corpora are available on request.

59 Especially in dialogues, such expressions can be discontinuous. These cases are also annotated in ARRAU.

60 `https://www.isi.edu/~marcu/discourse/Corpora.html`.

61 `http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/`.

62 `http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm`.

ARRAU contains about 350,000 tokens in total. There are 1,633 instances of non-NA anaphora: 631 from RST, 862 from TRAINS, 73 from GNOME, and 67 from PEAR. Even though RST is by far the largest subcorpus, TRAINS contains considerably more instances of non-NA anaphora, showing that non-NA anaphora is especially frequent in spoken data.

They identify and annotate these cases as follows. When a coder specifies that a referring expression is discourse-old, they ask the coder whether its antecedent was introduced using a phrase or a discourse segment. If the coder selects segment as the type of antecedent, they have to mark a sequence of (predefined) clauses as the antecedent. For the subset of non-NA annotations, Uryupina et al. do not report inter-annotator agreement.

All NPs have been manually annotated for a variety of features, including agreement features (gender, number, person), grammatical function (e.g., subj, obj, adjunct, np-mod) and semantic type (person, animate, concrete, organization, space, time, plan, numerical, abstract, unknown) and genericity.

The annotation of ARRAU is thoroughly documented in Uryupina et al. (2018) and the GNOME annotation guidelines.[63] The ARRAU corpus is publicly available from the LDC.[64] The ARRAU corpus has recently been used for the shared task *Resolution of discourse deixis*[65] (Poesio et al. 2018), organized by the NAACL Workshop on Computational Models of Reference, Anaphora, and Coreference (CRAC).

4.2.10 Lapshinova-Koltunski, Hardmeier, and Krielke (2018): The ParCorFull corpus. Lapshinova-Koltunski, Hardmeier, and Krielke present ParCorFull, a large corpus of parallel texts in English and German, annotated with coreference information. The corpus is intended both as a resource for NLP applications and as a basis for contrastive linguistic research in translation studies. It combines data from three sources: the test sets of two shared tasks (IWSLT 2013 and DiscoMT 2015), composed of TED talks, and news texts selected from the test set of another shared task (WMT 2017). The annotations are partly based on the ParCor corpus (Guillou et al. 2014).[66]

The English part of the corpus contains 82,379 tokens, the German part 78,350. In the English subset, there are 468 instances of non-NA anaphora, and in the German subset, there are 444.

The annotations cover NPs and different kinds of pronouns (personal, demonstrative, relative, and reflexive). In addition, pronominal adverbs are also included in the German subset. Such adverbs are fusions of locative adverbs *da, hier, wo* ('there, here, where') and certain prepositions, as in *damit* (literally 'therewith'). They occur rather often in German but are rarely considered in annotation projects. The annotation scheme covers a vast array of anaphora and coreference-related phenomena, including *one* anaphora, ellipsis, and comparative reference (repeated mentions as in, e.g., *the same students*). The annotation scheme is based on other coreference guidelines designed for multilingual data. The annotation covers non-NA anaphora, and allows for non-NAs in the form of VPs or (sets of) clauses. With regard to non-NA anaphora, the guidelines (Lapshinova-Koltunski and Hardmeier 2018) are not very detailed, however.

63 http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/.
64 https://catalog.ldc.upenn.edu/LDC2013T22.
65 http://anawiki.essex.ac.uk/dali/crac18/crac18_shared_task.html.
66 http://opus.nlpl.eu/ParCor/.

The corpus has been annotated by well-trained expert annotators. Lapshinova-Koltunski, Hardmeier, and Krielke (2018) compute inter-annotator agreement on two TED talks, using the mention overlap and entity-based CEAF scores (Luo 2005), and treating one of the annotators as the system output and the other as the reference ("truth"). This results for English in F_1-scores of 80.71% (mentions) and 74.13% ($CEAF_e$), and, for German, 76.54% (mentions) and 65.88% ($CEAF_e$), so that English appears to be easier to annotate than German.

ParCorFull is freely available,[67] and the repository includes the guidelines.

4.2.11 Further Corpora. We conclude this section by briefly describing further corpora with non-NA annotations, for languages other than English.

Kučová and Hajičová (2004), Nedoluzhko and Mírovský (2012): Prague Dependency Treebank. Nedoluzhko and Mírovský describe the coreference annotation of the Czech Prague Dependency Treebank,[68] which extends the original annotation of Kučová and Hajičová (2004). The Prague Dependency Treebank consists of 49,431 annotated sentences and about 1.8 million words in total. Coreference is annotated on the tectogrammatical layer, a kind of dependency structure of content words in which the meaning of functional words such as determiners or auxiliary verbs has been integrated into the content nodes. The coreference annotation includes all forms of anaphors (NPs, personal and demonstrative pronouns, etc.). Non-NA and NA anaphora are both marked as *coref_text* if their antecedent cannot be specified by grammatical means in the same sentence. The antecedent (NP, VP, clause, or sentence) is identified in terms of a (sub)tree, whose ID is used to annotate the anaphor. Multi-sentence antecedents are not explicitly marked. In this case the anaphor is just assigned the feature *segm*. The annotation uses a *principle of maximal size of an anaphoric expression*, which means that it always includes the whole subtree of an anaphor and antecedent, respectively. This approach avoids many consistency issues that would arise otherwise during annotation. It is difficult to extract statistics of non-NA anaphora from Kučová and Hajičová (2004) and Nedoluzhko and Mírovský (2012) because they do not differentiate between non-NA anaphora and NA anaphora.

A related resource is the parallel Prague Czech-English Dependency Treebank,[69] which consists of the one-million token English Wall Street Journal corpus of the Penn Treebank and its translation into Czech. Parts of the English coreference annotation was derived from the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein 2005)[70] whereas the Czech annotation was done from scratch (Nedoluzhko et al. 2016).

Recasens and Martí (2010): AnCora. The corpus consists of two data sets, one in Catalan (AnCora-CA) and one in Spanish (AnCora-ES), each with about 500,000 words.[71] The corpus is richly annotated: Besides lemma, part of speech, and syntactic information, the annotation includes various kinds of semantic information—for example, argument structures, thematic roles, semantic verb classes, and WordNet nominal senses. For coreference, all NPs are considered and marked as referential or non-referential.

67 https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2614.
68 https://ufal.mff.cuni.cz/pdt3.0.
69 http://ufal.mff.cuni.cz/pcedt2.0-coref.
70 https://catalog.ldc.upenn.edu/LDC2005T33.
71 http://clic.ub.edu/corpus/en.

Coreference links cover a range of phenomena, including split antecedents and non-NA anaphora (which they call discourse deixis). They distinguish between three types of non-NA relations: anaphoric reference to the same event-token, to the same event-type, and to the proposition.[72]

Navarretta and Olsen (2008). Navarretta and Olsen studied the equivalents of *this*, *that*, and *it* in Danish (455 instances) and Italian (114 instances) written and spoken data.[73] Their goal was to understand the use of these pronouns as a basis for their automatic processing. They annotated the following properties for each instance: the type of the pronoun, the antecedent, the semantic and syntactic type of the antecedent, and the distance between the anaphor and its antecedent in terms of clauses.

Dipper and Zinsmeister (2012). The authors annotated 225 instances of non-NA anaphora in German from the Europarl corpus (Koehn 2005)[74], concentrating on personal and demonstrative pronoun anaphors.[75] Their idea was to design an annotation procedure that makes the way speakers conceptualize the entities denoted by non-NAs explicit. For that, they suggested linguistic tests that help the annotators to identify the antecedent and determine the semantic types of the anaphors and antecedents.

Simonjetz and Roussel (2016). A related project examined the crosslingual behavior of shell nouns, involving the annotation of about 2,140 shell noun instances from the Europarl corpus in English and German. To increase annotation consistency, they concentrated on annotating candidates from a set of 50 predefined shell noun lemmas. In the subcorpus used to calculate inter-annotator agreement, the annotators agreed on the shell noun status of 1,140 of 1,329 instances, and for approximately 65% of the instances that were marked as shell nouns, identical antecedent spans were annotated (the antecedent spans overlap 96% of the time). Besides single non-NAs, they also considered multiple antecedents, nominalized antecedents, and instances of plural shell nouns.

4.3 Summary and Discussion

In this section, we started with the challenges associated with annotating non-NA anaphora. The primary challenge in annotating non-NAs is identifying their precise boundaries, which causes low inter-annotator agreement. Next, we described prominent resources for non-NA anaphora in detail, focusing on the goals of the annotations, the anaphoric expressions considered in annotation, the annotation schemes, and inter-annotator agreement.

Early annotation efforts such as Schiffman (1985), Passonneau (1989), and Botley and McEnery (2001) focused on verifying claims from the linguistics literature and observing trends in the usage of *this*, *that*, and *it*. Eckert and Strube (2000) and Byron (2003) annotated data in order to implement and test their rule-based resolution systems. Later approaches (e.g., Müller 2008; Kolhatkar and Hirst 2012; Kolhatkar, Zinsmeister, and Hirst 2013a,b; Uryupina et al. 2018) annotated corpora suitable for building supervised machine learning resolution systems.

72 AnCora Coreference guidelines: `http://clic.ub.edu/corpus/webfm_send/15`.
73 `https://cst.dk/dad/`.
74 `http://www.statmt.org/europarl/`.
75 The data is available from `https://github.com/kvarada/non-NA_Resources`.

Another distinction among these corpora is with respect to the domain, which affects the difficulty level of the annotation. If we assume a spectrum with closed-domain corpora at one end and open-domain corpora at the other, Byron's (2003) corpus and parts of Uryupina et al.'s (2018) ARRAU corpus would be close to the closed-domain end, as TRAINS93 dialogues and the GNOME corpus are closed domain corpora, where the topics or objects of discussion are fixed. On the other end of the spectrum, we have Eckert and Strube's (2000) annotation of a subset of the Switchboard corpus and Müller's (2008) annotation of unrestricted multi-party dialogues from the ICSI Meeting corpus. Kolhatkar and Hirst's (2012) corpus of MEDLINE abstracts is more varied than the TRAINS93 dialogue and the GNOME corpus, but it is still on the closed-domain side. The corpora of Kolhatkar, Zinsmeister, and Hirst (2013a,b) fall somewhere in the middle of the spectrum.

Another difference concerns the coverage of the annotation. We see three kinds of approaches to annotate non-NAs: annotating semantic type, annotating representative verbal head antecedents (which act as proxies for clausal linguistic antecedents), and annotating linguistic antecedents.

Annotating semantic type. Some approaches annotate semantic types, like event or fact. As we discussed in Section 3.2.1, the semantic type of the anaphor is determined by the anaphor's context. However, the semantic type of the antecedent, as determined by the meaning of the antecedent, also plays a role. The two types can differ, a phenomenon referred to as **referent coercion**.

In a first attempt, Gundel, Hedberg, and Zacharski (2004) annotated the semantic types of antecedents. They assumed that clauses denote eventualities (either events or states of affairs/situations, depending on whether the predicate is eventive or stative), and VPs denote either activities or states, so determining the antecedent's type was relatively straightforward in these cases. For the anaphor, semantic constraints of the predicate had to be used as cues. This proved difficult for multiple reasons: Either the predicates of the pronouns did not force a single interpretation, there was no suitable term to label the type, or the referent was too vague. They switched to annotating the kind of relation as direct or indirect instead.

Byron (2003) also annotated the anaphor's semantic type. In the TRAINS corpus, semantic types are strongly domain-dependent (e.g., a plan or the time taken by an action). The BUR corpus shows a greater variety of types, with types like events or processes. She also tried to label the relation between the antecedent and the anaphor, but the student annotators had too many problems with this task, and so it was abandoned.

Annotating verbal head antecedents. Müller's (2008) scheme (as well as Weischedel et al. [2013] and the Prague Dependency Treebank [Nedoluzhko et al. 2016]) marks representative verbal heads for non-NAs, assuming that they act as proxies for clausal or sentential antecedents. This scheme thus provides a degree of flexibility and is able to avoid some problems associated with annotating precise antecedents. However, there are two problems with this scheme. First, the verb gives only partial information about the antecedent and its type. Only marking a verb as the antecedent would not tell us whether we are talking about an event, a concept, or a fact. Moreover, if it is an event, for instance, it is not clear which arguments of the verb should be included in the antecedent. Second, antecedents with multiple verbs or with discontinuous antecedents cannot be expressed effectively with this annotation scheme.

Annotating linguistic antecedents. Eckert and Strube (2000), Byron (2003), Artstein and Poesio (2006), Kolhatkar and Hirst (2012) and Kolhatkar, Zinsmeister, and Hirst (2013a) mark clausal, sentential, and verbal syntactic constituents. The main issue is the underspecification of such antecedents—all references do not need to be fully specified for successful communication. Recasens (2008) suggests that computational approaches should bear this in mind and that annotation efforts must not insist on setting fixed boundaries in every case.

Several researchers point out the difficulties associated with annotating different aspects of this phenomenon, in particular with respect to identifying the precise boundaries of non-NAs. There is no standard way to report inter-annotator agreement for this kind of annotation. Some studies use Krippendorff's α with distance metrics such as Dice and Jaccard; others use Krippendorff's unitizing alpha. The agreement numbers in either case are rather low, especially for open domains such as newswire. Some studies report only observed percentage agreement with results in the range of about 0.40–0.55 Vieira et al. (2002); Dipper et al. (2011).[76]

Table 4 summarizes prominent annotation efforts in non-NA anaphora. The primary focus of annotation has been on the demonstratives *this* and *that* and the personal pronoun *it*. Most of the studies were carried out as preliminary investigations, and very few corpora are available for reuse. Also, the size of most of the corpora is relatively small for training a machine learning system. In Table 4, we mark publicly available corpora with an asterisk (*).

The data format and the tool used in the annotation process often have an impact on the design decisions of the annotation schemes or the workflow. The most commonly used annotation tools in non-NA anaphora annotation are: MMAX2 (Strube and Müller 2003), the AnCoraPipe annotation suite (Bertran et al. 2008), TrEd ("TreeEditor") (Pajas and Štěpánek 2008),[77] and PALinkA (Orăsan 2003). See Poesio et al. (2016) for a review of the currently available corpora for anaphora and tools to create such corpora.[78]

Although a few projects have attempted to annotate non-NA anaphora in a way that can be useful for the development of computational systems (e.g., the ASN Corpus by Kolhatkar, Zinsmeister, and Hirst and the ARRAU corpus by Uryupina et al.), if we want to see real progress in computational methods for this phenomenon, we will need larger, systematically annotated corpora for benchmarking computational systems. We have created a GitHub repository[79] for documenting all the relevant resources for non-NA anaphora.

5. Computational Approaches to Non-NA Anaphora

Attempts to appropriately interpret non-NA anaphora date back to the earliest days of natural language processing. Bobrow's (1964) high-school algebra problem solving system (STUDENT) was designed to handle such cases as in Example (41). Here, *this product* is interpreted as referring to the result of the preceding sentence, which is

76 The variation in the results is mainly due to the variation in the number of annotators, types of anaphors, and language of the corpora.

77 `https://ufal.mff.cuni.cz/tred/`.

78 See also Coref Annotator, which has recently been developed: `https://nilsreiter.github.io/CorefAnnotator/`.

79 `https://github.com/kvarada/non-NA_Resources`.

Table 4
Corpora in English for non-NA anaphora. For each work, we provide information about the number and type of non-NA instances in the anaphora column. "(+)" and "(‡)" mark corpora that indicate the antecedent or its semantic type, respectively. Publicly available corpora are marked with an asterisk "(*)"

Work	Corpus data	Anaphora
Schiffman (1985)	Transcribed career-counseling interviews	298 pronouns (*it*: 65, *that*: 233)
Webber (1991)	Essays, reviews, technical reports	96 pronouns (*it*: 15, *this*: 62, *that*: 19)
Eckert and Strube (2000)	Switchboard corpus (telephone conversations)	(+) 154 pronouns (*it*: 47, *this, that*: 107)
Byron (2003)	*TRAINS93 (task-oriented dialogues), *BUR (read news stories)	(+)(‡) 68 pronouns (*it*: 16; demonstratives: 52)
Poesio and Modjeska (2002, 2005)	GNOME (museum descriptions and pharmaceutical leaflets)	19 demonstratives
Botley and McEnery (2001); Botley (2006)	Associated Press, Hansard, and American Printing House for the Blind	403 demonstratives (*this*: 149, *that*: 244, *these*: 9, *those*: 1)
Gundel, Hedberg, and Zacharski (2002)	Santa Barbara Corpus of Spoken American English (spontaneous conversation)	(‡) 110 personal pronouns (*it*)
Artstein and Poesio (2006)	TRAINS91 (task-oriented dialogs)	(+) 28 instances (*it*: 2, demonstratives: *this*: 4, *that*: 20, *those*: 2) (experiment 1)
Hedberg, Gundel, and Zacharski (2007)	New York Times	(+)(‡) 178 pronouns[1] (*it, this, that*)
Pradhan et al. (2007)	OntoNotes (mix of genres)	(+)[2] 502 pronouns (*it*: 146, *this*: 85, *that*: 271)
Müller (2008)	ICSI meeting corpus (multi-party discussions)	(+)[2] 150 pronouns (*it, this, that*)
Kolhatkar and Hirst (2012)	**This issue* corpus (MEDLINE abstracts)	(+) 183 *this issue*
Kolhatkar, Zinsmeister, and Hirst (2013a); Kolhatkar (2015)	*ASN and *CSN corpora (New York Times corpus)	(+) 1,810 anaphoric instances (ASN), (+) 114,700 cataphoric instances (CSN) of six shell nouns
Uryupina et al. (2018)	*ARRAU (mix of genres)	(+) 1,633 pronouns and shell nouns
Lapshinova-Koltunski, Hardmeier, and Krielke (2018)	*TED talks, news	(+) 468 instances (pronouns, nominalizations, possibly shell nouns)

[1]These are cases marked as indirect by both annotators. Reference to events cannot be distinguished from NA anaphora in their scheme, cf. Section 2.1.5.
[2]Pradhan et al. (2007) and Müller (2008) only mark the antecedent's head verb.

syntactically represented with a non-NA (albeit not referring to an abstract entity in our sense).

(41) **A number is multiplied by 6**. **This product** is increased by 44. The result is 68. Find the number.

Winograd's (1972) SHRDLU system resolves *it* and *that* by remembering the last possible event. Furthermore, Winograd's heuristic stipulates that *it* refers to the event mentioned by the speaker, whereas *that* can refer to the last event mentioned by anyone.

Though they provide interesting clues, these heuristics do not cover the entire range of the phenomenon and they are unlikely to be particularly effective outside of the considerably restricted domains for which they were conceived. Since then, more sophisticated approaches have been developed that attempt to meet the challenges particular to this task.

What makes the resolution of non-NA anaphora particularly challenging with respect to conventional anaphora resolution or coreference resolution is that precisely those features that make those tasks tractable problems are the ones that are missing in this domain. Thus, systems that attempt the resolution of non-NA anaphora must attempt to access semantic or discourse-related information, which is not always easily accessible, in order to make resolution decisions.

Whereas NP coreference algorithms can easily and efficiently select an appropriate set of candidate antecedents, namely, by considering NPs only, this is not possible for non-NA anaphora. The antecedents can have a great variety of syntactic shapes and can be difficult to distinguish from one another, as in Example (2), repeated here. Here, *whether* may or may not be included in the antecedent and it is unclear to what degree these are different antecedents. Furthermore, the same constituent may represent various semantic entities, partly depending on which anaphor is used to refer to it and on the anaphor's context. It is unclear which of these potential candidates ought to be offered to an algorithm and when, since it would be inefficient to consider all their possible variations proactively.

(2) The municipal council had to decide **whether to balance the budget by raising revenue or cutting spending**. The council had to come to a resolution by the end of the month. **This issue** was dividing communities across the country.

Second, agreement features, such as number and gender, critical to the resolution of nominal anaphora, are generally absent for non-NA anaphora. Rather, the features that are useful for determining the compatibility of anaphors and their non-NAs tend to refer to levels of annotation, such as semantic and discourse structures, that are not easily accessible or generally available.

Finally, existing NP coreference algorithms can also generally depend on there being multiple references to a single entity, in which case each mention offers additional information about the entity being described, which in turn can be useful to resolution algorithms. Though some resources, such as ARRAU (Poesio et al. 2013),[80] include such referential chains, many others do not. However, even where these chains are available, their usefulness for non-NA anaphora is limited by the lack of agreement features and the ability of anaphors to adjust the types of their antecedents (cf. referent coercion in Section 3.2.1, Example (22)). As a result, resolution algorithms for non-NA anaphora generally must consider each instance in isolation.

80 See Section 4.2.9.

In this section, we will present some attempts that have been made to address these particular aspects of non-NA anaphora resolution and discuss their effectiveness as well as the implications this has for future work in this area.

5.1 Rule-Based Methods

Whereas the methods used by the historical systems previously mentioned relied on simple heuristics and tightly controlled domains to resolve non-NA anaphora, the systems and approaches described in the following involve more sophisticated algorithms. They incorporate linguistic knowledge about discourse structure and anaphoric reference and offer more detailed evaluations of their performance on naturally occurring data in a variety of domains.

5.1.1 Resolution as Linear Search. The approach of Eckert and Strube (2000) is intended to cover both NA anaphora (which they call individual anaphora), and non-NA anaphora (which they call discourse deixis), in the Switchboard corpus (Godfrey and Holliman 1993).[81]

To this end they use a simple discourse model consisting of two lists: individuals (i.e., referents of NA anaphors) are recorded in a list, the S(alience)-list, and ordered according to salience, whereas abstract discourse entities that have been referred to, for instance by non-NA demonstratives, are recorded in the A(bstract)-list. These two lists are incrementally updated as a text is processed.

Vendler (1967) and Asher (1993) have previously observed that the context of an anaphoric expression provides valuable information as to the nature of its referent, and by extension the range of possible antecedents. Eckert and Strube (2000) use the similar concept of *A(bstract)-* or *I(ndividual)-incompatibility* to determine which anaphoric expressions may or may not refer to non-NAs. As mentioned in Section 3.2.1, Eckert and Strube describe a particular anaphor instance as I-incompatible if its context does not allow for reference to an individual or concrete entity. I-incompatible contexts include sentences such as *x is true* or *x is correct*. A-incompatibility refers to contexts in which the anaphor may not refer to abstract objects.

In order to determine the non-nominal antecedent, Eckert and Strube (2000) suggest a *context ranking* algorithm, which is inspired by Webber's (1991) right-frontier constraint (see Section 3.2.4). The authors work with two units: *Initiation* and *Acknowledgment*. Initiations are the dialogue acts that convey semantic content, whereas Acknowledgments do not convey semantic content but have the pragmatic function of signaling that the other participant's utterance has been heard or understood. A single Initiation and the Acknowledgment that follows (if present) jointly constitute a **synchronizing unit**. The context ranking algorithm effectively carries out a linear search. It first searches for an appropriate antecedent in the A-list and, if there is none, it considers clauses in the same and then in previous Initiations. If an antecedent is found, the algorithm stops and adds the antecedent to the A-list. The algorithm determines a clause or sentence as the antecedent if it is:

(i) in the A-list (with all abstract objects previously referred to anaphorically in the same synchronizing unit—usually empty),

(ii) the clause to the left of the clause containing the anaphor in the same Initiation,

(iii) the rightmost main clause and subordinating clause to its right in the previous Initiation,

81 https://catalog.ldc.upenn.edu/ldc97s62.

(iv) the rightmost main clause in some previous Initiation.

After each synchronizing unit, all antecedents are cleared from the A-list.

Example (42), from Eckert and Strube (2000), illustrates the algorithm. The demonstrative *that* is resolved to the rightmost main clause and the subordinating clause to its right in the previous Initiation A.50 (rule (iii)) because the A-list is empty and there is no clause to the left of *that* in the same Initiation A.52.

(42)	**Initiation**	A.50	because if you tell everybody everything, **everybody in the world would know because they'd put it on TV**
	Acknowledgment	B.51	Right.
	Initiation	A.52	and **that** wouldn't do us any good. (SW3241)

Eckert and Strube (2000) evaluate their algorithm using three selected Switchboard dialogues, which were annotated by two annotators, as described in Section 4.2.2. Their algorithm had a precision of 63.6% and a recall of 70.0% for the non-NA anaphors in the data. According to the authors, the results show that the algorithm primarily has trouble with the classification of anaphors: If an anaphor was resolved incorrectly, it was usually also classified incorrectly, for example, a non-NA anaphor was incorrectly classified as a NA (individual) anaphor.

5.1.2 PHORA. Byron (2002, 2004) describes the design and evaluation of the PHORA algorithm for the resolution of pronominal reference to both NA and non-NA anaphora. Her algorithm constructs a discourse model, as a text is interpreted phrase by phrase. The discourse model involves two lists, one of *mentioned* entities and one of *activated* entities.

All referential NPs (i.e., excluding expletive *it*) result in discourse entities in the list of mentioned entities. One of these is considered the focus. Though Byron's algorithm leaves the exact method of calculating the focus open, in the implementation that was evaluated, the left-most NP in each clause was considered the focus. Constituents that can act as the antecedents of non-NA anaphors (in PHORA: infinitives, gerunds, entire sentences, subordinate clauses) result in discourse entity (DE) *proxies* in the list of activated entities. Whereas mentioned discourse entities remain in the model for the duration of the discourse, activated entities only remain in the model for the duration of the following clause. This behavior reflects Webber's (1988) right-frontier constraint. However, once an abstract entity has been referred to anaphorically, it is treated as an ordinary mentioned entity, thus enabling multiple references to a single abstract entity.

Similar to the approach of Eckert and Strube (2000), the compatibility of a pronoun with a candidate antecedent is determined according to their semantic types, which are inferred from their context. For example, verbs impose semantic restrictions on their arguments: In the sentence *Load them into the boxcar*, the semantics of *load* requires its theme argument to be of the semantic type cargo. Certain copular adjectives have a similar function. In the statement *that's correct*, the entity which is described as being correct must be a proposition.

The algorithm also provides for a set of *referring functions* that govern the conversion of discourse entities between types and from proxies to abstract entities. Depending on the entity's syntactic, semantic, and speech act–relevant properties, the entity may be coerced to a certain semantic type as required (cf. the process of referent coercion in Section 3.2.1). For instance, the referring function *Proposition()* can be applied to a sentential proxy *d* in an assertion and coerces the proxy to its propositional content

Table 5
Referring functions adapted from Byron (2002, page 84). A referring function takes a proxy *d* as its input and coerces it into an appropriate referent.

Referring function	Constraints on the proxy *d*	Example proxy
Situation(*d*)	Sentence with tensed stative verb	**The train is red.**
Event(*d*)	Sentence with tensed eventive verb	**It gets there late.**
$\text{Kind}_A(d)$/ $\text{Kind}_E(d)$	Infinitive form of action/event	**To load them** takes an hour.
	Gerund form of action/event	**Loading them** takes an hour.
Proposition(*d*)	Each assertion or yes/no–question	**I think that he's an alien.**
	that sentence	I think that **he's an alien**.
	if /*when* subordinating clause	If **he's an alien...**

Table 6
Example discourse model adapted from Byron (2002, page 85).

Constituent	Number	Semantic class	Specificity	Referent	Salience
Engine 1	Sing	ENGINE	Indiv	ENG1	focus
Avon	Sing	CITY	Indiv	AVON	mentioned
the oranges	Plural	ORANGE	Indiv	ORANGES1	mentioned
to get [the] oranges	Sing	Functional	Kind	proxy	activated
Sentence (43a)	Sing	Functional	Indiv	proxy	activated

characterized as a proposition. Table 5 contains a number of such referring functions for different semantic types.

When a demonstrative pronoun, such as *that*, is encountered, the algorithm first attempts to resolve it to one of the activated entities from the previous sentence. This can only be successful if the type of the pronoun (as determined by its predication context) is compatible with that of a potential antecedent. For example, the discourse model after reading the sentence of Example (43a) is shown in Table 6 (from Byron 2002). At that point, the discourse model contains several concrete entities (ENG1, AVON, ORANGES1) as well as two abstract entities, whose semantics is not spelled out; instead, the constituent itself serves as a proxy of its semantics (*to get [the] oranges, Engine 1 goes to Avon to get [the] oranges*).

(43) a. **Engine 1 goes to Avon to get the oranges.**
b. **That** takes two hours.

Here the demonstrative *that* is resolved as follows. First the predicate complements are examined and checked in the list of predicates. The semantics of *take (time)* requires an argument of the type event. Following the search order for demonstratives, the algorithm first looks for the activated DEs, using the referring function *Event(d)*. The function will be successful on the proxy DE *Engine 1 goes to Avon to get the oranges* in (43a), as the verb *goes* is a tensed eventive verb. Thus the referring function states that *that* refers to the event of Engine 1 getting to Avon, which takes two hours.

In contrast to Eckert and Strube (2000), who present a system design that was not implemented, Byron (2002, 2004) presents an implemented system. She evaluates her algorithm using 180 pronoun instances from the TRAINS93 corpus (Allen and Heeman 1995),[82] a subset of the dialogues annotated in the Byron (2003) study

82 `http://www.cs.rochester.edu/research/cisd/resources/trains.html`.

described in Section 4.2.4. The algorithm is compared against a baseline that does not distinguish between personal and demonstrative pronouns and uses salience alone to select antecedents. This baseline resolved 37% of pronouns in the evaluation set correctly. With all system components active, the PHORA system resolved 72% of the pronouns correctly. Byron notes that many of the remaining errors are the result of the linear discourse structure used in the system, which could potentially be improved by detecting embedded dialogues and structural shifts.

5.1.3 A Hybrid Centering Theory–Based Discourse Model. Pappuswamy, Jordan, and VanLehn (2005) describe an algorithm that is intended to recognize and resolve non-NA anaphors in computer-mediated tutorial dialogues on physics. The algorithm implements a discourse model that combines elements of Centering Theory (Grosz, Weinstein, and Joshi 1995) of Grosz and Sidner's (1986) theory of discourse structure, and focus stacks (Pfleger, Alexandersson, and Becker 2003) (see Section 3.2.3).

The main idea of this discourse model is that a given discourse has a focus, which remains the same for a few sentences before shifting to a new entity. The entity that is the focus is normally pronominalized at that point in the discourse; therefore, keeping track of the focus should be useful in the interpretation of pronominal expressions. Pappuswamy, Jordan, and VanLehn (2005) make use of two types of focus structures in their algorithm, a global focus and a local focus. The global focus stack keeps track of topics as they relate a discourse as a whole, and each member of this stack has its own local focus stack, which keeps track of discourse objects (i.e., potential antecedents) pertaining to this topic.

Once the algorithm has established that a particular mention of *this*, *that*, or *it* requires a non-NA, it will first attempt to replace the pronoun with the previous sentence. If the substitution is "complete and coherent," then the previous sentence is accepted as the antecedent; otherwise, in the case of *that* and *it*, the list of utterances in the same discourse segment is searched for a similarly compatible substitution. In the case of *this*, the algorithm will also check if the discourse center is compatible with the global or local focus, in which case the global or local focus is taken to be the antecedent. Pappuswamy, Jordan, and VanLehn (2005) do not specify how completeness of the substitutions is determined. Coherence is determined based on constraints from Centering Theory and from domain knowledge about the topic.

Pappuswamy, Jordan, and VanLehn (2005) did not implement the algorithm. They tested it by hand on 40 referring expressions in their corpus of physics tutorial dialogues. They report that their algorithm resolved 91% of the discourse-deictic anaphors (20 out of 22 cases) successfully. Interestingly, whereas all of the *that* and *it* instances were resolved successfully, only 75% of the *this* instances (for which a slightly different algorithm was used, involving the discourse center and global/local focus) were resolved successfully, suggesting that this strategy is less effective than the one used for *that* and *it*.

5.2 Machine Learning Approaches

5.2.1 Resolving this, that, *and* it. The earliest attempt to resolve pronominal anaphors to non-NAs using machine learning methods was that of Strube and Müller (2003). This approach aimed to resolve anaphoric reference to both NA and non-NA antecedents. Every NP that was not an indefinite NP was considered a potential anaphor, and every NP that preceded a potentially anaphoric NP was considered a potential antecedent, insofar as it was compatible in terms of agreement features. All such pairs were used to train the system to resolve references to nominal antecedents. When the potential

anaphor was an instance of *it* or *that*, however, candidate antecedents were generated by selecting S and VP constituents from the two sentences preceding the anaphor. In order to approximate the right-frontier constraint (Webber 1991), constituents that were not the last constituent in a given sentence were not considered candidates, because they were considered "inaccessible" by the algorithm.[83]

The feature set used in Strube and Müller (2003) is split between agreement-based features that are only intended for NP-coreference and those that are useful for non-NA anaphora. (Table 7 later in this article provides a summary of the features used by Strube and Müller (2003) as well as those used by the other systems described in this section.) This includes features that record a verb's preference for nominal or non-nominal arguments, which were inspired by Eckert and Strube's (2000) and Byron's (2002) observations regarding the preferences established by a pronoun's predicative context. These preferences are implemented here as a list of verbs together with the frequencies with which they were used with arguments of the NP, VP, or S types. The authors also use a feature intended to capture the importance of a particular antecedent candidate with respect to the dialogue as a whole. This feature is implemented as TF*IDF, comparing word frequencies in specific documents with their frequencies in the complete set of Switchboard dialogues (Godfrey and Holliman 1993). For non-NAs, an average TF*IDF value was calculated based on all of the words in the antecedent. Strube and Müller (2003) use a decision tree classifier to decide between the potential antecedents for each annotated anaphoric instance. When tested with all pronouns, both those with nominal and those with non-nominal antecedents, the system receives an F_1-score of 47.42, 56.74 precision, and 40.72 recall. Resolution performance for third-person neuter pronouns, the only pronouns with non-NAs, was $F_1 = 19.26$, $P = 40.41$, and $R = 12.64$.

Müller (2007, 2008) describes an algorithm for the resolution of instances of *this*, *that*, and *it* in a corpus of spoken dialogues (described in Section 4.2.6). Instances of these pronouns are resolved either to NPs or VPs; in the case of VPs, only the verbal head is annotated, such that the verbal head substitutes equally for VP or S antecedents. Candidate antecedents are those that occur within a given temporal distance: 9 seconds for NPs and 7 seconds for VPs. However, VP candidate antecedents are only generated for instances of *that* or objects of the verb *do*. Müller uses these data to train a logistic regression classifier, which decides whether or not a particular anaphor–antecedent pair constitutes a case of anaphoric reference. This approach uses an interesting means of estimating the I-incompatibility (cf. Section 5.1.1) of lemmas using corpus frequency counts. The likelihood that an adjective can be used to modify an abstract entity is calculated as the conditional probability of the adjective to occur with a *to*-infinitive complement, as in Equation (1). A similar formula estimates the probability of an adjective to occur with a *that*-sentence complement, which similarly is indicative of abstract entities.

$$\frac{\#it('s \mid is \mid was \mid were)\ \text{ADJ}\ to}{\#it\ ('s \mid is \mid was \mid were)\ \text{ADJ}} \tag{1}$$

Similar features encode the likelihood of verbs to appear with sentence complements and the semantic compatibility of anaphors and NP antecedents. The system's best

83 Note the problematic nature of this interpretation of the right-frontier constraint: The constraint was originally intended to apply to *discourse* structure, whereas this algorithm, and several that follow its lead, apply the constraint to syntactic parse trees.

performing configuration receives only an F_1-score of 12.59 (with $P = 13.43$ and $R = 11.84$) for non-NAs. The author notes, however, that the contribution of the corpus-based probability estimates did not significantly improve the system's overall performance. (Nevertheless, similar features do appear in later systems.) Despite the system's low performance, it is the first to attempt the *fully-automatic* resolution of non-NA anaphora: Previous systems relied either on the anaphors being pre-selected for resolution or on costly domain-specific knowledge and manual annotations.

Chen, Su, and Tan (2010) resolve instances of *it*, *this*, and *that* in the OntoNotes Release 2.0 data set (Hovy et al. 2006) to their verbal antecedents using a ranking support vector machine (SVM) model with a composite kernel. The composite kernel combines a series of positional, lexical, and syntactic features with the output of a convolution tree kernel, which encodes the similarity between two syntactic structures and allows the system to better distinguish antecedent candidates.

Their approach considers the preceding verbs in the anaphor's sentence, together with the verbs from the previous two sentences as the candidate antecedent set. All of the antecedents in this set are compared pairwise using the SVM model, and the antecedent that wins the greatest number of such comparisons is selected as the anaphor's antecedent, provided that this antecedent's score exceeds a certain threshold. If no antecedent meets this criterion, then the anaphor instance is taken to refer to a nominal antecedent and left unresolved.

The authors collect two types of negative instances: those antecedents that belong to the candidate set of non-NA anaphors and antecedents from the candidate set of NA anaphors. The intuition is that both types of negative antecedent provide different types of information about the form of non-NAs. However, because this results in an even greater imbalance between positive and negative instances than would ordinarily be present, the authors also use random desampling to balance the training data. Chen, Su, and Tan (2010) evaluate their system using 10-fold cross validation, and their results show that using additional negative instances and balancing the training data both improve performance (primarily by increasing recall). In its best performing configuration, their system has an F_1-score of 57.9 ($P = 62.6$, $R = 54.0$).

Jauhar et al. (2015) build on the approach of Müller (2007), separating the task into two discrete stages: classification and resolution. The classification stage involves a decision as to whether or not a particular instance of *this*, *that*, or *it* refers to a non-NA and is thus a candidate for resolution. If the instance is classified as a positive instance, then in the resolution stage, potential antecedent candidates are considered and a second classifier decides for each antecedent whether or not an anaphoric link is to be established. For both stages, maximum entropy classifiers are used. As in Müller (2007), features that estimate the likelihood of a verb to have a clausal or verbal argument or which approximate I-incompatibility using corpus frequency counts are included (here, from the in-house Google News corpus[84]), as well as similar features measuring the association strength between a pronoun's parent verb—as determined by dependency parsing—and an antecedent's verbal head, among others. The system's overall performance is 22.2 F_1-score ($P = 22.6$, $R = 21.8$), as compared with a baseline of $F_1 = 16.5$ ($P = 15.3$, $R = 17.9$). The system performs best for the pronoun *that* ($F_1 = 28.0$, compared with $F_1 = 17.1$ for *this*), and the pronoun *it* proves especially hard to resolve ($F_1 = 2.4$).

84 The data set is not freely available.

5.2.2 Resolving Shell Nouns. The first study to examine shell nouns (Section 3.1.2) in a computational context, Kolhatkar and Hirst (2012), addressed anaphoric shell noun instances of *this issue* in the MEDLINE abstracts corpus (the annotations are described in Section 4.2.7). In particular, they built a candidate ranking model to rank all eligible antecedent candidates of *this issue* in the corpus. The eligible candidates are extracted from the same sentence or the two previous sentences whose type (according to the Penn Treebank tag set) is contained in the set {S, SBAR, NP, SQ, SBARQ}. They also considered infinitive phrases as eligible candidates, which are typically analyzed by automatic parsers such as (S (VP *to go to school*)). The set of candidate constituents is then expanded (in part to attenuate the effects of parser errors) by also adding "mixed-type constituents," which are created by concatenating NP and VP sister constituents, as shown in Example (44). Here, the correct antecedent is of mixed type; it consists of a NP constituent and its sister VP constituent while the PP part of the parent S node is not part of the antecedent.

(44) (S (PP Given these data) (, ,) (NP **decreasing HTD to < or = 5 years**) (VP **may have a detrimental effect on patients with locally advanced prostate cancer**) (. .)) Only a randomized trial will conclusively clarify **this issue**.

The candidates are encoded using a series of automatically extracted features (see Table 7 later in this article) and then ranked using SVMs. The system's top-ranked candidate matched the manually annotated candidate in 60.78% of the cases in the system's best-performing configuration, which corresponds to an F_1-score of 77.92 in terms of token overlaps in system and gold-standard antecedents.

Whereas Kolhatkar and Hirst (2012) used relatively small amounts of manually annotated training data, Kolhatkar, Zinsmeister, and Hirst (2013b) attempted to circumvent this training data bottleneck by extracting training data from instances conforming to structurally determined relations of shell noun phrases and their shell content (see Table 3).[85] An illustration is shown in Example (45), where the shell content (*whether animal testing is cruel*) given in a copula structure can be easily extracted automatically.

(45) Of course, **the central, and probably insoluble, issue** is **whether animal testing is cruel**. (NYT)

Kolhatkar, Zinsmeister, and Hirst hypothesize that the shell content of these structurally determined relations bear some degree of similarity to the antecedents of anaphoric shell noun instances. For example, specific syntactic patterns are associated with specific shell nouns (e.g., *whether* clauses are typically used to express questions and issues). They exploit this insight and use the relatively easy-to-gather information to automatically resolve the harder anaphoric cases. They gather automatically labeled training data using typical shell noun patterns from Table 3 and train an SVM ranker with this training data. They apply these trained models to predict antecedents of harder anaphoric cases.

The authors evaluated the effectiveness of this approach using crowdsourced annotations: Participants were asked to select the best antecedent candidate for a given anaphoric shell noun instance from among the top ten randomly ordered alternatives

85 See Section 4.2.8.

provided by the system.[86] The annotators agreed with the system's top-ranked candidate in between 35% (in the case of *decision*) and 72% (*reason*) of the tested instances.

Kolhatkar and Hirst (2014) focus, in contrast to the previous two studies, primarily on improving the resolution of cataphora-like shell noun instances. Though the relative reliability of lexico-syntactic patterns was one of the motivations for the approach in Kolhatkar, Zinsmeister, and Hirst (2013b), there remain a number of unclear cases, which the Kolhatkar and Hirst (2014) study attempted to address. The reason for these unclear cases is that shell nouns take different types of one or more semantic arguments, and the problem is identifying the appropriate semantic argument that is the shell content of the shell noun phrase in the given context. For instance, in Examples (46) and (47), the shell nouns are resolved to the postnominal *that* clause and the copula *that* clause, respectively.[87] Resolving the shell noun phrase *the usual reason* in Example (47) involves identifying (a) that *reason* generally expects two semantic arguments: cause and effect, (b) that the cause argument (and not the effect argument) represents the shell content, and (c) that a particular constituent in the given context is the cause argument.

(46) **The fact [that a major label hadn't been at liberty to exploit and repackage the material on CD]**$^{\text{general factual content}}$ meant that prices on the vintage LP market were soaring. (NYT)

(47) Although there are many technical objections, **the usual reason** [why courts have rejected DNA tests that seem to show guilt]$^{\text{effect}}$ is [**that scientists disagree about how to calculate the odds that there is a match between cells from a suspect and cells from a crime scene**]$^{\text{cause}}$. (NYT)

The system implemented in this study integrates a number of preferences, as described in Schmid's (2000) study on the linguistic properties of shell nouns. In order to evaluate their system, the authors compiled a gold standard annotation using crowdsourcing and compared it with a baseline configuration that uses lexico-syntactic patterns (as in Table 3) alone to make resolution decisions. Where the baseline averages 57% accuracy, the system manages either 64% or 69%, depending on whether or not the linguistic clues described by Schmid are used or not, respectively. These clues, such as that *reason* disprefers the N-clause pattern, are a significant help for nouns such as *reason* (+24%) and *fact* (+13%), because these nouns tend to occur in clearly defined syntactic environments. For others, such as *difficulty* or *problem*, the clues do not improve the system's performance.

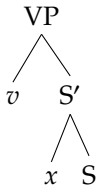
5.2.3 A Deep Learning Approach to Non-NA Anaphora Resolution. Marasović et al. (2017) describe an attempt to resolve both pronouns and shell nouns that refer to non-NAs using an LSTM-based (Long Short-Term Memory) mention-ranking model and an innovative method to generate training data. Because of the requirement of neural methods for large amounts of training data, the authors generate training data by automatically pairing the context of anaphor instances with antecedent constituents.

For instance, whenever a verb (v) occurs with an embedded sentential constituent S', as in Example (48), the S' node represents an artificial antecedent, and is replaced in the original sentence by an appropriate (yet randomly selected) anaphor, either *this*, *that*, or *it*. The extracted artificial antecedent (49a) is then paired with the resulting sentence

86 See Section 4.2.8 for a more in-depth description of the annotation process.

87 Note that the postnominal *that* clause in Example (46) is not a relative clause: The fact in question is not an argument of *exploit and repackage*.

with the anaphor (49b) as a training instance (Examples (48) and (49) are from Marasović et al. 2017).

(48) a. 

b. He doubts [$_{S'}$ [$_{S}$ a Bismarckian super state will emerge that would dominate Europe], but warns of "a risk of profound change in the [. . .] European Community from a Germany that is too strong, even if democratic"].

(49) a. [$_{S'}$[$_{S}$ **a Bismarckian super state will emerge that would dominate Europe**]]

b. He doubts **this**, but warns of "a risk of profound change in the [. . .] European Community from a Germany that is too strong, even if democratic".

The approach posits that there exists some as yet unclear semantic relation between the sentences in Example (49a) and (49b), which the neural network encodes. In order to encode this relation, the model uses a 'Siamese' bidirectional LSTM neural network architecture, whose name refers to the twinned LSTM components working in parallel: One of these components is applied to the anaphoric sentence and the other to the antecedent. Each of the representations thus derived are then combined into a single, joint representation of the anaphor–antecedent pair. The system's ultimate resolution decisions are then derived from this joint representation.

The authors test the performance of their model on the resolution of both shell nouns and pronominal non-NA anaphora in two settings. In the first setting, they consider shell nouns only and use the anaphoric and cataphoric-like shell noun data sets (ASN and CSN) from Kolhatkar, Zinsmeister, and Hirst (2013a,b) and Kolhatkar and Hirst (2014), described in Sections 4.2.8 and 5.2.2. CSN is used for training, ASN for evaluation. On the ASN data set from Kolhatkar, Zinsmeister, and Hirst (2013b), the authors report significantly improved results, ranking the correct antecedent first in 76.09% (for *decision*) to 93.14% (for *possibility*) of the instances (this measure can be thought of as roughly similar to accuracy).

In the second setting, they use the artificially generated data for training and evaluate their system using data from the ARRAU corpus (Poesio et al. 2013), described in Section 4.2.9, which contains both shell nouns and pronominal non-NA anaphors. In its best configuration, their model ranked the correct antecedent first in 29.06% of the pronominal instances and in 51.89% of the shell noun instances.

5.2.4 Relevant Features. Table 7 contains an overview of the features used in machine learning algorithms for the resolution of non-NA anaphora. Note that, as the table is intended to give a general picture of the features used in various systems, certain simplifications were necessary: Features that are highly domain-specific or do not appear to be sufficiently useful for the task (e.g., the features removed during feature selection by Jauhar et al. [2015]) are omitted here. The features included in the table have been sorted according to whether they primarily involve information about the anaphor, a candidate antecedent, or the relation between the two. Though the studies do not tend to implement many of these features in precisely the same way, we tried nevertheless to summarize them in a way that allows for useful comparisons across systems. Features relating to "parents" imply the use of dependency-parsed data and

Table 7
Comparison of features used for the resolution of non-NA anaphora. SM03 = Strube and Müller (2003); M07 = Müller (2007, 2008); CST10 = Chen, Su, and Tan (2010); KH12 = Kolhatkar and Hirst (2012); JGGR15 = Jauhar et al. (2015).

	SM03	M07	CST10	KH12	JGGR15
Anaphor features					
Form of anaphor	✓	✓		✓	✓
Is demonstrative		✓			✓
Grammatical function	✓				
Embedding depth	✓	✓			
Position in sentence					✓
Verb presence					✓
Parent lemma					✓
Parent transitivity					✓
Probability of parent to govern a clause	✓				✓
Parent + grammatical function					✓
Dependency label path to root					✓
Complement clause type		✓			
Co-occurance patterns			✓		
Antecedent features					
Grammatical function of antecedent	✓		✓	✓	
Length in tokens		✓		✓	
Number of arguments		✓			
Syntactic type		✓		✓	
Semantic role				✓	
Contains modal verb		✓		✓	
Contains finite verb		✓			
Contains subordinating conjunction		✓		✓	
Sentence w/existential *there* construction		✓			
Probability of progressive/perfect aspect		✓			
Embedding depth	✓	✓	✓	✓	
Dependency label path to root					✓
Syntactic parse tree			✓		
Is negated					✓
Is transitive		✓			✓
Right-frontier constraint	(✓)	✓			✓
I-incompatibility		✓			✓
Lexical overlap w/title	(✓)			✓	
Lexical overlap w/anaphor sentence				✓	
Neighbor is preposition or punctuation				✓	
Relational features					
Distance in tokens	✓	✓	✓		✓
Distance in sentences	✓	✓	✓	✓	✓
Distance (other units)	✓	✓			
Anaphora vs. cataphora		✓		✓	✓
Whether antecedent is ancestor of anaphor		✓			✓
In same sentence		✓		✓	
Whether grammatical functions match	✓				
Whether tense of ante. and ana. parent match		✓			

refer to this dependency relation: The "parent lemma" is the lemma of the word on which an anaphor depends. "Parent transitivity" accordingly describes whether or not the anaphor's parent, insofar as it is a verb, is a transitive verb. Some of the rows contain check marks in parentheses: These are intended to show that, though a particular feature plays a role in a system's functioning, it is implemented in a significantly different way from the other systems. The first such set of parentheses reflect the fact that Strube and Müller (2003) use an interpretation of the right-frontier constraint to disallow certain antecedent candidates rather than as a classification feature. The second set of parentheses refers to the authors' use of TF*IDF as opposed to actual lexical overlap in order to judge a candidate's relation to the text's overall topic.

5.3 Summary

Most of the early attempts to resolve non-NA anaphora are limited to particular domains, and their rule-based design means that these approaches may not easily generalize to other domains. They are largely dependent on rich linguistic information, which is difficult to acquire, because it requires considerable time and expertise to produce, and it is unclear to what degree this information can be gathered automatically. (Though one possibility would involve leveraging existing semantic resources, such as WordNet [Fellbaum 1998], as Eckert and Strube [2000] and Byron [2002] suggest.) Perhaps owing to these difficulties, these algorithms have not generally been implemented computationally—namely, annotation as well as evaluation for these algorithms has usually been carried out manually.

Later attempts to solve this problem make use of machine learning methods. They use surface-based features and information that is readily available and easy to gather automatically. However, knowledge-poor methods do not tend to be particularly effective, as is evidenced by relatively low recall in general. The implemented methods for the resolution of shell nouns appear to be somewhat more effective, due either to additional information provided by the shell noun or to the restricted syntactic environment shell nouns prefer. Most recently, computational approaches have leveraged deep learning techniques to get at the latent information in the relation between an anaphor's context and its antecedent. This has led to the development of interesting means of data generation and encouraging results.

It is interesting to note the generally high performance of the rule-based approaches with respect to statistical methods. It would be interesting to see to what degree future algorithms could integrate insights from these older methods with the more sophisticated machine learning approaches of today. An overview of the resolution algorithms just described can be found in Table 8.

6. Discussion and Conclusion

This survey gives an overview of the state of the art of resolving non-NA anaphora in English. We reviewed the main aspects of the relevant rule-based and machine-learning approaches, including those using deep learning techniques. As a basis for this, we described major annotation efforts in the domain and provided an extensive list of existing resources for English. To equip the reader with a better understanding of the phenomenon's intricacies, we detailed the linguistic properties of both of the elements involved in non-NA anaphora: the pronominal anaphor or shell noun and its non-nominal antecedent.

Table 8
Resolution approaches covering non-NA anaphora. P = precision, R = recall, F = F_1-measure, A = accuracy, CE = candidate extraction, CS = candidate selection.

Work	Anaphoric expressions	Approach		Performance
		CE	CS	
Eckert and Strube (2000)	47 instances of *it* and 107 instances of demonstratives from the Switchboard corpus	Previously referred discourse entities, new referents created with coercion	Linear search with context ranking	P = 63.6, R = 70.0
Byron (2002, 2004)	180 test pronouns from TRAINS93 problem-solving dialogues	Discourse entities and proxy discourse entities	Referring functions	A = 72 (all pronouns, baseline = 37)
Strube and Müller (2003)	20-fold cross validation on 553 dialogues from the Switchboard corpus	All S and VP constituents from the last two valid sentences	Decision tree with subcategorization and TF*IDF features	F = 19.26, P = 40.41, R = 12.64 (3rd person neuter pronouns)
Pappuswamy, Jordan, and VanLehn (2005)	40 instances of *this*, *that*, *it* from Why-2 corpus of physics tutoring dialogues	All utterances and sentences within the same discourse segment	Hybrid approach based on Centering Theory and morphological, syntactic, semantic, and topical constraints and preferences	A = 91
Müller (2007, 2008)	343 anaphoric chains of *it*, *this*, *that* from ICSI meeting corpus	NP and VP chunks in discourse order	Logistic regression classifier	F = 12.59, P = 13.43, R = 11.84 (VP antecedents)
Chen, Su, and Tan (2010)	502 instances of *it*, *this*, *that* from OntoNotes 2.0	Preceding verbs from current sentence and all verbs from previous two sentences	Ranking SVM with composite kernel	F = 57.9, P = 62.6, R = 54.0
Kolhatkar and Hirst (2012)	183 instances of *this issue* from MEDLINE abstracts	All syntactic constituents from the current and two preceding sentences	Lexical, semantic, syntactic features and candidate ranking SVM	Exact match, 60.78%; partial match, 77.92%
Kolhatkar, Zinsmeister, and Hirst (2013b)	1,810 instances of anaphoric shell nouns (ASN) from the New York Times corpus	All syntactic constituents from automatically created training data	SVM ranking with automatically labelled training data and syntactic, semantic, lexical features	A = 35 to 72, depending upon the shell noun
Jauhar et al. (2015)	2,075 instances of *this*, *that*, and *it* from the CoNLL-2012 corpus		Logistic regression classifier	F = 22.2, P = 22.6, R = 21.8
Marasović et al. (2017)	ASN + CSN, ARRAU data: 397 shell nouns, 203 pronouns		Mention-ranking model with a LSTM–Siamese Net architecture	ASN: A = 76.09 to 93.14, depending on the shell noun; ARRAU: A = 29.06 (pronouns), A = 51.89 (shell nouns)

This survey could not cover all aspects related to the phenomenon. Most notably, we only scratched the surface of the semantics and pragmatics of referring to event-like referents (e.g., Asher 1993; Webber et al. 2003). Secondly, for practical reasons, we restricted this survey mostly to approaches dealing with the English language. As with many natural language processing tasks, there are some universal aspects that will generalize to the resolution of non-NA anaphora in other languages, such as using shell nouns as anaphors in addition to pronouns. In addition, there are language-specific aspects that can be learned from the training data, for example, the realization preferences of antecedents. But there are also language-specific properties that will require new approaches, such as zero anaphors in Italian, in which case there is no overt anaphoric element (Navarretta 2007).

There are a number of open questions that have yet to be addressed. We conclude with a list of some possible research avenues.

Identifying the antecedent string. As noted before, non-NAs are underspecified, and it is not always clear what the ground truth antecedent for a given anaphor instance is, which poses a serious challenge to computational systems. Current computational systems do not make the fine-grained distinctions between antecedent strings implied by the use of particular pronouns. If these systems were in a position to make these distinctions, it could result in considerable performance improvements. An important question related to underspecification is whether identifying the precise text segments representing non-NAs is useful in NLP applications or not. We believe that this question does not have a straightforward answer, and the answer will be primarily task-dependent. For instance, if we are building a dialogue system and we want to tackle examples like Example (3), it will be enough to identify the possible semantic types for the demonstrative pronoun *that*. On the other hand, if we are building an assistant for ESL learners, it could be helpful to mark the antecedent strings in the text. Investigating the usefulness of non-NA anaphora resolution systems in NLP applications will be a concrete step towards making progress in this field.

Another open question related to identifying antecedent strings is related to marking discontinuous strings as antecedents. A property of non-NAs that is not modeled well in current approaches is that the antecedent is not necessarily a continuous string of words that is contained in a single sentence (see Example (5b)). Some antecedents are discontinuous strings, which exclude modal verbs and negation markers, for example, and, furthermore, some of them comprise more than one sentence.

Combining data-driven and knowledge-driven approaches. Early approaches to resolve non-NA anaphora exploit linguistic properties discussed in Section 3. That said, these approaches are largely dependent on rich linguistic annotations, which are difficult to gather automatically, especially for different domains. Current approaches to non-NA anaphora resolution do not rely on domain-specific rules but work in a fully data-driven manner, exploiting the current state of the art of NLP preprocessing tools and standard lexical resources, so that they can be applied to raw text input. Carrying out systematic error analyses of resolution systems in order to understand what kind of patterns the data-driven systems are able to capture and combining the ideas from both knowledge-driven and data-driven approaches (e.g., by modeling the predication context of anaphors and antecedents with dense vector representations) will be promising research avenues.

Identifying instances of non-NA anaphora. Most of the approaches described in this survey are optimized for the resolution of non-NA anaphora, and they ignore the task of resolving nominal antecedents or disambiguating between instances of nominal and non-NAs (notable exceptions are, e.g., Müller 2008, Chen, Su, and Tan 2010, and Jauhar et al. 2015). If the resolution of non-NA anaphora is integrated into a general coreference resolution system, it will require some disambiguation between anaphors that expect a non-NA and those that do not. In addition, systems are needed that reliably disambiguate between referential and non-referential uses of pronouns as well as pronouns referring to the extra-linguistic context of an utterance.

Identifying the context size for the antecedent. It seems advantageous to split the task of identifying the antecedent into two steps (Kolhatkar and Hirst 2012); first, by specifying the target sentence, and second, by identifying the actual string within this sentence. If the target sentence is predefined, current approaches are able to identify the verbal head or to approximate the antecedent string with a reasonable degree of accuracy. That said, the first step of identifying the target sentence is more difficult, and is not fully solved yet. One practical issue here is that the number of antecedent candidates rises steeply with each additional context sentence to be considered, and it is very likely that more than one candidate string would be a good fit. Current systems are able to discriminate between the preferences for different anaphor types to a certain extent, but it would also be helpful to impose independent textually motivated restrictions, such as paragraph structure, or constraints related to discourse structure.

Modeling modality and text registers. This is related to the previous question. Many linguistic phenomena are dependent on modality (e.g., written vs. spoken), and also on text register or genre (for the latter, see, e.g., Webber [2009] for discourse connectives, and Kunz and Lapshinova-Koltunski [2015] for cohesive devises in general, including coreference). It is an open question whether systems will improve when resolution is conditioned on these textual meta-properties.

Incorporating predication context in the resolution systems. We saw that predication context of the anaphor provides clues about the semantic type of the antecedent (Webber 1988; Eckert and Strube 2000; Byron 2004). That said, from a computational linguistics perspective, it is not clear to what extent these semantic types (e.g., *fact*-type antecedent or *situation*-type antecedent) help in automatic resolution of non-NA anaphora. Resources such as FrameNet (Baker, Fillmore, and Lowe 1998) include subcategorization frames for nouns such as *issue* and *fact*. However, there are no resources that describe characteristic properties associated with these semantic types, which would help a computational system to identify the text segment representing those semantic types.

Moreover, it is not clear what level of granularity of semantic types is appropriate to describe the antecedent. For instance, in Example (22), repeated here as Example (50), the antecedent can be described as a proposition (highly abstract level) or an event (more granular). The question is what level of granularity suffices in choosing the right antecedent among other competing candidates.

(50) **John crashed the car.**

a. His girlfriend couldn't believe **it**. (proposition)

b. **It** happened yesterday at 10 in the morning. (event)

c. **This** shows how careless he is. (fact)

Additional contextual effects. We will close this survey by discussing an example that is out of the range of today's resolution systems.

In Example (51), adapted from the New York Times corpus (Sandhaus 2008), until one knows the agent of the approving event, it is not clear whether the antecedent of *this decision* is allowing the sale of standing-room-only tickets for adults or selling the standing-room-only tickets for adults. When the agent is the marketing department, as in (51a), we immediately know that the decision is about selling because the function of marketing departments is usually related to selling and not regulating; whereas when the agent is the safety regulators as in (51b), the decision will be about whether or not something is to be allowed. Option (51c) is ambiguous and both antecedents could work in that context.

(51) In principle, he said, airlines should be allowed to sell standing-room-only tickets for adults — as long as **this decision** was approved by
   a. their marketing departments.
   b. the safety regulators.
   c. the right people.

Although the semantic type of the antecedent stays the same in all three readings, the antecedent's interpretation will change, as well as the antecedent string. This is interesting because the resolution will only be determined by the agent of the approving event after the shell noun phrase itself has been uttered/processed. It demonstrates how we sometimes need common sense knowledge to identify the appropriate antecedent referent when several candidates are available in the vicinity of the anaphor. It is an open question whether such subtle context dependencies can be captured by data-driven approaches of anaphora resolution.

Acknowledgments

We sincerely thank the reviewers and editors for their thoughtful, constructive, and valuable comments on our manuscript. We also thank Graeme Hirst, Bonnie Webber, and Massimo Poesio for their valuable suggestions regarding the terminology used in this article. The research reported here was partially funded by Ruhr-Universität Bochum and Universität Hamburg.

References

Afantenos, Stergos D. and Nicholas Asher. 2010. Testing SDRT's right frontier. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1–9, Beijing.

Ahn, David, Valentin Jijkoun, Gilad Mishne, Karin Müller, Maarten de Rijke, and Stefan Schlobach. 2004. Using Wikipedia at the TREC QA track. In *Proceedings of the Thirteenth Text Retrieval Conference TREC*, Gaithersburg, MD.

Allen, James and Peter Heeman. 1995. TRAINS Spoken Dialog Corpus LDC95S25. Linguistic Data Consortium, Philadelphia, PA.

Anand, Pranav and Daniel Hardt. 2016. Antecedent selection for sluicing: Structure and content. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1243, Austin, TX.

Artstein, Ron and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *brandial'06: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)*, pages 56–63, Potsdam.

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Asher, Nicholas. 2008. Troubles on the right frontier. *Pragmatics and Beyond New Series*, 172:29–52.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, volume 1, pages 86–90, Montreal.

Bergsma, Shane and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Proceedings of the 8th Discourse Anaphora*

and Anaphor Resolution Colloquium, pages 12–23, Faro.
Bertran, Manuel, Oriol Borrega, Marta Recasens, and Bàrbara Soriano. 2008. AnCoraPipe: A tool for multilevel annotation. *Procesamiento del Lenguaje Natural*, 41:291–292.
Bobrow, Daniel G. 1964. A question-answering system for high school algebra word problems. In *Proceedings of AFIPS Fall Joint Computer Conference*, 26, pages 591–614, San Francisco, CA.
Botley, Simon and Tony McEnery. 2001. Demonstratives in English: A corpus-based study. *Journal of English Linguistics*, 29(1):7–33.
Botley, Simon Philip. 2006. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.
Boyd, Adriane, Whitney Gegg-Harrison, and Donna Byron. 2005. Identifying non-referential *It*: A machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, MI.
Byron, Donna K. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87, Philadelphia, PA.
Byron, Donna K. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. Technical report, Computer Science Department, University of Rochester, Rochester, NY.
Byron, Donna K. 2004. *Resolving pronominal reference to abstract entities*. Ph.D. thesis, University of Rochester, Rochester, NY.
Böhmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank. In Anne Abeillé editor, *Treebanks: Building and Using Parsed Corpora*, Dordrecht, Netherlands, pages 103–127.
Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, volume 16, pages 85–112, Aalborg.
Chafe, Wallace L., editor . 1980. *The Pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Ablex, Norwood, NJ.
Chen, Bin, Jian Su, and Chew Lim Tan. 2010. A twin-candidate based approach for event pronoun resolution using composite kernel. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 188–196, Beijing.
Clark, Herbert H. 1975. Bridging. In *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing (TINLAP)*, pages 169–174,Cambridge, MA.
Consten, Manfred, Mareile Knees, and Monika Schwarz-Friesel. 2007. The function of complex anaphors in texts: Evidence from corpus studies and ontological considerations. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in text: Cognitive, formal and applied approaches to anaphoric reference*, number 86 in Studies in Language Companion Series. John Benjamins, Amsterdam, Netherlands, pages 81–102.
Cornish, Francis. 1992. So be it: The discourse-semantic roles of *so* and *it*. *Journal of Semantics*, 9(2):163–178.
Cornish, Francis. 2007. English demonstratives: Discourse deixis and anaphora. A discourse-pragmatic account. In R. A. Nilson, N. A. A. Amfo, and K. Borthen, editors, *Interpreting Utterances: Pragmatics and Its Interfaces. Essays in Honour of Thorstein Fretheim*. Novus Press, pages 147–166.
Dhillon, Rajdip, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide, Technical report, ICSI TR-04-002. International Computer Science Institute (ICSI), Berkeley, CA.
Dipper, Stefanie, Christine Rieger, Melanie Seiss, and Heike Zinsmeister. 2011. Abstract anaphors in German and English. In *Anaphora Processing and Applications*, volume 7099 of *Lecture Notes in Computer Science*, pages 96–107, Springer.
Dipper, Stefanie and Heike Zinsmeister. 2012. Annotating abstract anaphora. *Language Resources and Evaluation*, 46(1):37–52.
Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, and Nii Martey. 2000–2005. Santa Barbara Corpus of Spoken American English, parts 1-4. Linguistic Data Consortium, Philadelphia, PA.
Eckert, Miriam and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.
Fellbaum, Christiane. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Feng, Vanessa Wei and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, MD.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):387–382.

Fort, Karën, Adeline Nazarenko, and Sophie Rosset. 2012. Modeling the complexity of manual annotation tasks: A grid of analysis. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 895–910, Mumbai.

Francis, Gill. 1986. *Anaphoric Nouns*. Number 11 in Discourse Analysis Monographs. University of Birmingham (ELR), Birmingham, UK.

Francis, Gill. 1994. Labelling discourse: An aspect of nominal group lexical cohesion. In M. Coulthard, editor, *Advances in Written Text Analysis*, Routledge, London, pages 83–101.

Fraurud, Kari. 1992. *Processing Noun Phrases in Natural Discourse*. Ph.D. thesis, Stockholm, Institutionen för lingvistik.

Godfrey, John and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Linguistic Data Consortium, Philadelphia, PA.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara J., Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Guillou, Liane, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie L. Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 3191–3198, Reykjavik.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1988. On the generation and interpretation of demonstrative expressions. In *Proceedings of the 12th Conference on Computational Linguistics*, volume 1, pages 216–221, Budapest, Hungary.

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

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 2002. Pronouns without explicit antecedents: How do we know when a pronoun is referential? In *Proceedings of the 4th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC)*, Lisbon.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 2004. Demonstrative pronouns in natural discourse. In *Proceedings of the 5th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC)*, pages 81–86, São Miguel.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 2005. Pronouns without NP antecedents: How do we know when a pronoun is referential? In A. Branco, T. McEnory, and R. Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive and Computational Modeling*. John Benjamins, pages 351–364.

Gundel, Jeanette K., Michael Hegarty, and Kaja Borthen. 2003. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12(3):281–299.

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge, London.

Hardmeier, Christian, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon.

Hedberg, Nancy, Jeanette K. Gundel, and Ron Zacharski. 2007. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of the 6th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC)*, pages 31–36, Lagos.

Hegarty, Michael, Jeanette K. Gundel, and Kaja Borthen. 2001. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27(2–3):163–186.

Hirst, Graeme. 1981. *Anaphora in Natural Language Understanding: A Survey*, volume 119 of *Lecture Notes in Computer Science*. Springer.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, NY.

Huddleston, Rodney D. and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Ivanič, Roz. 1991. Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *International Review of Applied Linguistics in Language Teaching*, 29(2):93–114.

Jauhar, Sujay Kumar, Raul D. Guerra, Edgar Gonzàlez, and Marta Recasens. 2015. Resolving discourse-deictic pronouns: A two-stage approach to do *it*. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 299–308, Denver, CO.

Joty, Shafiq, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia.

Kamp, Hans. 1979. Events, Instants and Temporal Reference. In Rainer Bäuerle, Urs Egli, and Christoph Schwarze, editors, *Semantics from different points of view*, Springer, Berlin, Heidelberg, New York, pages 376–417.

Karttunen, Lauri. 1976. Discourse referents. In J. D. McCawley, editor, *Syntax and Semantics 7: Notes from the Linguistic Underground*. Academic Press, New York, NY, pages 363–385.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, 5:79–86.

Kolhatkar, Varada. 2015. Resolving Shell Nouns. Ph.D. thesis, University of Toronto.

Kolhatkar, Varada and Graeme Hirst. 2012. Resolving "this-issue" anaphora. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pages 1255–1265, Jeju Island.

Kolhatkar, Varada and Graeme Hirst. 2014. Resolving shell nouns. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 499–510, Doha.

Kolhatkar, Varada, Heike Zinsmeister, and Graeme Hirst. 2013a. Annotating anaphoric shell nouns with their antecedents. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 112–121, Sofia.

Kolhatkar, Varada, Heike Zinsmeister, and Graeme Hirst. 2013b. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 300–310, Seattle, WA.

Krippendorff, Klaus. 1995. On the reliability of unitizing contiguous data. In Peter V. Marsden, editor, *Sociological Methodology*, volume 25. Blackwell, Cambridge, MA, pages 47–76.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*, second edition. Sage, Thousand Oaks, CA.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*, third edition. Sage, Thousand Oaks, CA.

Kučová, Lucie and Eva Hajičová. 2004. Coreferential relations in the Prague dependency treebank. In *5th Discourse Anaphora and Anaphor Resolution Colloquium*, pages 97–102, Azores.

Kunz, Kerstin and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1):258–288.

Lakoff, Robin. 1974. Remarks on *this* and *that*. In *Papers from the Tenth Regional Meeting of the Chicago Linguistics Society*, pages 345–356, Chicago, IL.

Lapshinova-Koltunski, Ekaterina and Christian Hardmeier. 2018. Coreference corpus annotation guidelines.

Lapshinova-Koltunski, Ekaterina, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: A parallel corpus annotated with full coreference. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki.

Le Nagard, Ronan and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala.

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

Loáiciga, Sharid, Liane Guillou, and Christian Hardmeier. 2017. What is it?

Disambiguating the different readings of the pronoun 'it'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1325–1331, Copenhagen.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Stroudsburg, PA.

Luperfoy, Susann. 1991. Discourse Pegs: A Computational Analysis of Context-dependent Referring Expressions. Ph.D. thesis, University of Texas at Austin, Austin, TX.

Lyons, John. 1977. *Semantics, II*. Cambridge University Press, Cambridge, UK.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marasović, Ana, Leo Born, Juri Opitz, and Anette Frank. 2017. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 221–232, Copenhagen.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miller, Philip. 2011. The choice between verbal anaphors in discourse. In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications*, Lecture Notes in Computer Science, Springer.

Mitkov, Ruslan. 2002. *Anaphora Resolution*. Routledge, London, UK.

Müller, Christoph. 2006. Automatic detection of nonreferential *it* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–56.

Müller, Christoph. 2007. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823, Prague.

Müller, Christoph. 2008. *Fully Automatic Resolution of* It, This *and* That *in Unrestricted Multi-Party Dialog*. Ph.D. thesis, Universität Tübingen, Tübingen, Germany.

Navarretta, Costanza. 2007. A contrastive analysis of abstract anaphora in Danish, English and Italian. In *Proceedings of the 6th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC)*, pages 103–109, Lagos.

Navarretta, Costanza and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2046–2052, Marrakesh.

Nedoluzhko, Anna and Ekaterina Lapshinova-Koltunski. 2016. Abstract coreference in a multilingual perspective: A view on Czech and German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, pages 47–52, San Diego, CA.

Nedoluzhko, Anna and Jiří Mírovský. 2012. Extended textual coreference and bridging relations in PDT 2.0. Technical report. Charles University, Prague.

Nedoluzhko, Anna, Michal Novák, Silvie Cinkova, Marie Mikulová, and Jiří Mírovskỳ. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris.

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

Orăsan, Constantin. 2003. PALinkA: A highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39–43, Sapporo.

Orăsan, Constantin. 2007. Pronominal anaphora resolution for text summarisation. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. pages 430–436, Borovets.

Ostendorf, Mari, Patti Price, and Stefanie Shattuck-Hufnagel. 1996. Boston University Radio Speech Corpus. Linguistic Data Consortium, LDC catalogue entry LDC96S36, Philadelphia, PA.

Pajas, Petr and Jan Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 673–680, Manchester, UK.

Palmer, Martha, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Pappuswamy, Umarani, Pamela W. Jordan, and Kurt VanLehn. 2005. Resolving

Discourse Deictic Anaphors in Tutorial Dialogues. In Claudia Sassen, Anton Benz, and Peter Kühnlein, editors. *Constraints in Discourse*, pages 95–102, Dortmund University, Germany.

Passonneau, Rebecca J. 1989. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 51–59, Vancouver.

Passonneau, Rebecca J. 2004. Computing Reliability for Coreference Annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1503–1506, Lisbon.

Pfleger, Norbert, Jan Alexandersson, and Tilman Becker. 2003. A robust and generic discourse model for multimodal dialogue. In *Proceedings of the 3rd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco, Mexico.

Poesio, Massimo. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of LREC*, pages 211–218, Athens.

Poesio, Massimo, Ron Artstein, Olga Uryupina, Kepa Rodriguez, Francesca Delogu, Antonella Bristot, and Janet Hitzeman. 2013. The ARRAU Corpus of Anaphoric Information LDC2013T22. Linguistic Data Consortium, Philadelphia, PA.

Poesio, Massimo, Yulia Grishina, Varada Kolhatkar, Nafise Sadat Moosavi, Ina Rösiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora Resolution with the ARRAU Corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, LA, USA.

Poesio, Massimo and Natalia N. Modjeska. 2002. The THIS-NPs hypothesis: A corpus-based investigation. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Conference (DAARC)*, pages 157–162, Lisbon.

Poesio, Massimo and Natalia N. Modjeska. 2005. Focus, activation, and *this*-noun phrases: An empirical study. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing*, volume 263. John Benjamins, pages 429–442.

Poesio, Massimo, Amrita Patel, and Barbara Di Eugenio. 2006. Discourse structure and anaphora in tutorial dialogues: An empirical analysis of two theories of the global focus. *Research on Language and Computation*, 4(2–3):229–257.

Poesio, Massimo, Simone Ponzetto, and Yannick Versley. 2010. Computational models of anaphora resolution: A survey. `https://www.researchgate.net/publication/265893367_Computational_Models_of_Anaphora_Resolution_A_Survey`. Unpublished manuscript.

Poesio, Massimo, Sameer Pradhan, Marta Recasens, Kepa Joseba Rodríguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In Massimo Poesio, Stuckardt Roland, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, Theory and Applications of Natural Language Processing. Springer, pages 97–140.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Poesio, Massimo, Roland Stuckardt, and Yannick Versley, editors. 2016. *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer, Berlin.

Poesio, Massimo, Patrick Sturt, Ron Artstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42(2):157–175.

Polanyi, Livia. 1985. A theory of discourse structure and discourse coherence. In *Proceedings of the 21st Meeting of the Chicago Linguistics Society*, pages 306–322, Chicago, IL.

Pradhan, Sameer S., Lance A. Ramshaw, Ralph M. Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the International Conference on Semantic Computing (ICSC)*, pages 446–453, Irvine, CA.

Quarteroni, Silvia. 2007. *Advanced Techniques for Personalized, Interactive Question Answering*. Ph.D. thesis, University of York, York, UK.

Recasens, Marta. 2008. Discourse deixis and coreference: Evidence from AnCora. In *Proceedings of the Second Workshop on Anaphora Resolution (WAR)*, pages 73–82, Bergen.

Recasens, Marta and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources & Evaluation*, 44(4):315–345.

Sandhaus, Evan. 2008. The New York Times Annotated Corpus. LDC catalogue entry LDC2008T19, Philadelphia, PA.
Schiffman, Rebecca J. 1985. *Discourse Constraints on 'it' and 'that': A Study of Language Use in Career-Counseling Interviews*. Ph.D. thesis, University of Chicago, Chicago, IL.
Schmid, Hans Jörg. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*, volume 34 of *Topics in English Linguistics*. De Gruyter, Berlin, Germany.
Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.
Siegel, Sidney and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, NY.
Simonjetz, Fabian and Adam Roussel. 2016. Crosslinguistic annotation of German and English shell noun complexes. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pages 265–278, Bochum.
Steinberger, Josef, Mijail Kabadjov, Massimo Poesio, and Olivia Sanchez-Graillet. 2005. Improving LSA-based summarization with anaphora resolution. In *Proceedings of the Human Language Technology Conference (HLT) and the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–8, Vancouver.
Strube, Michael and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo.
Taboada, Maite and Loreley Wiesemann. 2010. Subjects and topics in conversation. *Journal of Pragmatics*, 42(7):1816–1828.
Uryupina, Olga, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2018. Annotating a broad range of anaphoric phenomena, in a variety of genres: The ARRAU corpus. *Journal of Natural Language Engineering*. To appear.
Uryupina, Olga, Mijail Kabadjov, and Massimo Poesio. 2016. Detecting non-reference and non-anaphoricity. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Theory and Applications of Natural Language Processing*, Springer, Berlin, pages 385–409.
Vendler, Zeno. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.
Vendler, Zeno. 1968. *Adjectives and Nominalizations*. Mouton, The Hague, Netherlands.
Vicedo, José L. and Antonio Ferrández. 2008. Coreference in Q & A, In *Advances in Open Domain Question Answering*. Springer, Dordrecht, pages 71–96.
Vieira, Renata, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2002. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Conference (DAARC)*, pages 385–427, Lisbon.
Webber, Bonnie. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 674–682, Suntec.
Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.
Webber, Bonnie Lynn. 1979. *A Formal Approach to Discourse Anaphora*. Garland.
Webber, Bonnie Lynn. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122, Buffalo, NY.
Webber, Bonnie Lynn. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.
Weischedel, Ralph and Ada Brunstein. 2005. BBN Pronoun Coreference and Entity Type Corpus LDC2005T33. Linguistic Data Consortium, Philadelphia, PA.
Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA.
Winograd, Terry. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.
Winter, Eugene. 1977. A clause-relational approach to English texts: A study of some predictive lexical items in written discourse. *Instructional Science*, 6(1):1–92.