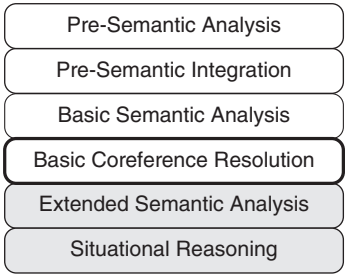
Linguistics for the Age of AI

The title-level DOI for this work is:

doi:10.7551/mitpress/13618.001.0001

5

Basic Coreference Resolution



The process of Basic Semantic Analysis described in the previous chapter attempts to disambiguate the words of input and establish their semantic dependencies. For some inputs, like *A squirrel is eating a nut*, this process will result in a complete, unambiguous meaning representation. But for most inputs, it leaves loose ends, such as residual ambiguity (multiple analyses are possible), incongruity (no analyses work out well), and underspecification (an imprecise analysis has been posited in anticipation of a more precise one).

This chapter considers one class of outstanding issues: underspecifications resulting from the need for textual coreference. *Textual coreference* refers to linking different mentions of the same entity in a so-called *chain of coreference* within a language input. For example, the constituents with matching subscripts in (5.1) are coreferential.

(5.1) I handed the card$_1$ to Nina$_2$. She$_2$ read it$_1$ silently and then aloud. (COCA)

Textual coreference is just one part of the much larger enterprise of *reference resolution*, which involves linking actual or implied mentions of objects and events to their anchors in an agent's memory. Agents undertake full reference resolution during Situational Reasoning (chapter 7), but they carry out many prerequisites to that during this stage of Basic Coreference Resolution. By contrast, most current NLP systems attempt only textual coreference—moreover, only select aspects of it (see section 1.6.8).

5.1 A Nontechnical Introduction to Reference Resolution

Since only linguists who already engage in the study of reference are likely to be familiar with the full problem space, we will start with an extended, example-based introduction. This is just a warm-up, intended to convey how much work needs to be done. It does not yet organize the phenomena or their treatment into a model—that will come later.

Some readers might find the introduction sufficient to satisfy their curiosity about the main topic of this chapter—textual coreference. The model of textual coreference is quite detailed, as it must cover many disparate phenomena; for the casual reader, the whole chapter beyond the introduction might read like a deep dive. There is no getting around this

complexity, but there is a choice: one can study, casually read, or briefly skim the description of the model. Any of these should suffice as preparation for later chapters.

5.1.1 Definitions

A *referring expression* (RefEx) is a word or phrase that has referential function: that is, it points to some object or event in the world. An entity that helps to resolve a RefEx (that is, to specify its meaning) is called its *sponsor*. In (5.2), he_{1A} is the sponsor for he_{1B}, and he_{1B} is the sponsor for $himself_{1C}$.

(5.2) As he_{1A} walks toward the exit, he_{1B} admires $himself_{1C}$ in the mirror behind the bar. (COCA)

We use the term *sponsor* (following Byron, 2004) rather than the more common *antecedent*, because not every sponsor is, in fact, an antecedent:

1. Whereas an antecedent must come before the RefEx in a speech stream or text, a sponsor can come either before or after it.
2. Whereas an antecedent must be part of the linguistic context, a sponsor can also be in the real-world context—as when something is pointed to and then visually perceived by the interlocutor.
3. Whereas an antecedent is assumed to have a strict coreference relationship with a RefEx, a sponsor can stand in various semantic relationships with it. For example, in (5.3) *a couple* is the sponsor of—though not coreferential with—*the husband*: it introduces two people into the context, only one of whom is coreferential with *the husband*.

(5.3) A voice came from across the dining room, where a couple was finishing their meal. "It happened to us too!" the husband volunteered. (COCA)

Entities that are evaluated as potential sponsors are called *candidate sponsors*. In (5.4), both *Einstein* and *Bohm* are candidate sponsors for the RefEx *he*.

(5.4) Einstein told Bohm that he had never seen quantum theory presented so clearly as in Bohm's new book … (COCA)

The *window of coreference* is the span of text/discourse where the sponsor is sought. Ideally, it is the most local segment of the discourse that is about the given topic. Stated differently, the window of coreference should not extend back into a portion of the discourse that is about something else. Unfortunately, it is currently beyond the state of the art to reliably, automatically analyze discourse structure. Therefore, most systems—including ours, for the moment—use a fixed window of coreference, usually a couple of sentences.

As mentioned earlier, coreference resolution involves linking textual elements that refer to the same thing. This is different from *reference* resolution, which involves linking the

meanings of text strings—or things perceived using other channels of perception, such as vision—to their *anchors* in a person's or agent's memory.[1] If there is no existing anchor—that is, if the entity or event is new to the agent—a new anchor must be created. Textual coreference is often a necessary step toward full reference resolution, but it is never the end stage. The end stage always involves modifying the LEIA's memory. This happens during Situational Reasoning.

Many *words* in language inputs do not refer and, therefore, are excluded from reference resolution procedures. These include such things as pleonastic *it* (*It is raining*) and non-compositional components of idioms (*He kicked the bucket*[2]). Basic Semantic Analysis, which is carried out before reference resolution, determines whether an entity is referential. Stated briefly:

- All referring expressions, and *only* referring expressions, end up as concept instances in the TMR for an input. In English, nouns (including pronouns) and verbs can be referring expressions.
- Only concept instances are subject to reference resolution.
- By specifying which concept instances comprise a TMR, Basic Semantic Analysis solves the problem of detecting which entities in texts are referring expressions, which is a well-known challenge for knowledge-lean approaches.
- The TMR does retain, as part of its metadata, a trace of the form of the referring expression in the input: for example, *a fox* versus *the fox* versus *this fox* versus *it*. These linguistic clues influence which coreference procedures are run.

5.1.2 An Example-Based Introduction

For readers new to this topic, the number of reference-oriented reasoning challenges that can pile up in a short text might come as a surprise.[3] Consider the following excerpt from a piece by Sabrina Tavernise called "Buying on Credit Is the Latest Rage in Russia" (*New York Times,* 2003).

> New advertisements are appearing on Moscow's streets and subways. Comic-book-style stories portray the new quandaries of the Russian middle class. "If we buy the car, we can't afford to remodel the apartment," says a woman with a knitted brow, in one ad. Then comes the happy ending. Her husband replies, smiling: "We can do both! If we don't have enough, we'll take a loan!"

To fully interpret this excerpt, a person or LEIA must understand the following.

1. The meaning of *new advertisements*, *comic-book-style stories*, *one ad*, *a loan*, and *a woman* are among the mentions of new entities that must generate new anchors in memory.

2. *A knitted brow* must generate a new anchor and must also be linked, using the PART-OF-OBJECT relation, to the anchor for the woman whose brow it is.
3. *Her husband* must generate a new anchor and must also be linked to the anchor for the woman, using the relation HAS-SPOUSE.
4. The interpretation of *we* involves combining into a set the anchors for the woman and her husband, which do not form a syntactic constituent (i.e., they appear in different portions of the text). Note that the first mention of *we* comes before the mention of the woman and her husband, thus requiring coreference resolution using a *post*cedent, not an *ante*cedent.
5. Understanding the meaning of *Moscow's streets and subways* requires linking the anchor for some unspecified set of streets and subways to the anchor for the city, Moscow.
6. The *Moscow* in question must be understood as the one in Russia. This can be inferred from the title or by reasoning that only the most well-known *Moscow* is likely to be written about. Both readers and intelligent agents (who could have access to very large lists of geographical place names), might realize that there are many towns and cities called *Moscow* worldwide.
7. Although noun phrases with *the*—which are called *definite descriptions* since they use the definite article—often refer to entities previously mentioned in the text, this is far from always the case. There are conditions under which *the* should not trigger the search for a textual sponsor; this occurs, for example,
    a. when the entity has restrictive postmodification: *the new quandaries of the Russian middle class* (*of the Russian middle class* is the postmodifier that licenses the use of *the* with *quandaries*);
    b. when the entity includes a proper name modifier: *the Russian middle class*;
    c. in clichés and idioms: *the happy ending*;
    d. when the meaning of the entity is generic: *the personal check*;
    e. when there is semantic ellipsis—that is, the omission of words/meanings that are necessary to fully understand the text but are not syntactically obligatory: *the car* and *the apartment* do not refer to just any car and any apartment—*the car* is the one the couple is thinking of buying and *the apartment* is the one they already own or rent.
8. All cases of ellipsis (i.e., missing elements) must be detected and resolved. In *If we don't have enough*, the meaning *money* must be understood as what is lacking. Moreover, that lacking must be attributed to the couple who needs the money to buy a car and remodel their apartment.

9. The meaning of all EVENTS—most often realized as verbs in text—must undergo reference resolution, just like OBJECTS. Events in this excerpt are conveyed by the words *appearing*, *portray*, *buy*, *afford*, *remodel*, *says*, *comes*, *replies*, *smiling*, *do*, *have*, *take*. Although most of these do not have a textual coreferent (instead, they directly create new anchors in memory), *do (both)* does have a textual coreferent: the set comprised of the event instances *buy (the car)* and *remodel (the apartment)*.
10. *The happy ending*, which must be a new anchor in memory, must be linked to this particular comic-book-style story, not stories in general. In addition, *the happy ending* must be understood as a cliché, which, in this context, conveys that the entire passage is a spoof.

Although this example does not exhaust or systematically organize all types of overt and elided RefExes, it is sufficient to illustrate the overall magnitude of the reference problem and the extent to which reference decisions must be integrated with overall semantic analysis and reasoning about language and the world.

5.1.3 A Dozen Challenges

As the next step in our big-picture overview, let us consider some issues that are more challenging than they might appear at first blush.[4]

Challenge 1. *Detecting RefExes can be difficult*. A given text string can be referential in some contexts and nonreferential in others, as shown by the contrastive examples in table 5.1.

Challenge 2. *The surface features (number, gender, animacy) of RefExes and their sponsors need not match, although they most commonly do*. For example, in English, ships can be referred to as *she*, and a beloved inanimate object can be referred to as *he* or *she*.

(5.5) Do not touch my car—she is only ever to be driven by me!

Likewise, *they* is making fast inroads as a gender-neutral singular pronoun.

(5.6) If someone got hurt, they would blame it on some outside force … (COCA)

Table 5.1
Referential and nonreferential uses of the same types of categories

RefEx type	Referential use	Nonreferential use
Definite description	Look at *the boat*!	On *the one hand*, …; on *the other hand*, ….[5]
Indefinite description	*A bug* just landed on your head.	Danny is *a plumber*.[6]
Pronoun	Take the vase and put *it* there.	*It* is not a good idea to lie.
Verb	George *has* a red car.	George *has*[7] finished painting the house.

Challenge 3. *Two noun phrases with the same head (the same main noun) do not necessarily corefer.* Although head-matching, like surface-feature matching, is a useful heuristic for identifying coreferential categories, it is quite normal for different entities of the same type, like the guards in (5.7), to be referred to by NPs with the same head noun.[8]

(5.7) The guard flipped a switch in his booth and a strip of spikes hinged up from the asphalt. The other guard set to circling the truck with the dog. (COCA)

Challenge 4. *Some personal pronouns can have a specific, a generalized, or a hybrid referent.* For example, in English the pronoun *you* can refer to one or more specific animate entities (5.8), people in general (5.9), or a nonspecific hybrid of those two meanings, implying "you and anyone else in the same position" (5.10).[9]

(5.8) If you eat another piece of pizza you'll explode.

(5.9) It's tough to live in the suburbs if you don't drive.

(5.10) If you speed you risk getting a ticket.

Similarly, *they* can refer to specific or nonspecific individuals, the latter illustrated by (5.11).

(5.11) They say it will rain tomorrow.

Challenge 5. *Fully interpreting referring expressions can require making implicatures.* For example, if a sixty-year-old married woman says, *We're going to Italy this summer*, the normal implication is that she and her husband are going. But if she were thirty, had kids, and was going to Disneyland, the kids would be implied as well.

Challenge 6. *Coreferential entities need not be of the same syntactic category.* For example: an NP can corefer with a verb (5.12), a predicate-adjective construction (5.13), a modality expressed by a verb (5.14), a set of NPs that must be dynamically combined (5.15), or a span of text (5.16).

(5.12) [Both expressions instantiate the concept INVADE]
If you have been invaded and that invasion is an accomplished fact several years down the line, you can not ignore it. (COCA)

(5.13) [All three expressions instantiate the concept AESTHETIC-ATTRIBUTE with a value of 1]
[Ramona:] And just remember how beautiful it is out here. It's really gorgeous. [Elam:] The beauty of the Flint Hills is especially apparent when you drive the ranch loop with Jane. (COCA)

(5.14) [Both expressions instantiate (MODALITY (TYPE EFFORT) (value 1))]
They had never said much to each other, she and her father, and when they tried now, their efforts felt strained and pointless. (COCA)

(5.15) [A set must be dynamically composed from *the Falcons* and *the Bills*]
The Falcons started with the Bills in Atlanta in Week 4. The two teams were tied at 17–17 in the fourth quarter, but the Bills pulled it out 23–17. (COCA)

(5.16) [The first clause, minus the discourse adverb *well*, corefers with *that*]
Well by the 1970s the risk factors has caused the Chesapeake Bay to loose [*sic*] 99 percent of its native oyster population. That's bad. (COCA)

Challenge 7. *Referentially related entities are not always coreferential*. Instead, they can express semantic relationships other than coreference, such as a set-member relationship (5.17), an instance-type relationship (5.18), different instances of a given type (5.19), and so-called bridging references,[10] in which the mention of one entity implicitly introduces another into the frame of discourse (5.20–5.23).[11]

(5.17) [Set-member relationship]
And I recently talked to a couple. And the wife, she was a nurse. (COCA)

(5.18) [Instance-type relationship]
The repairman canceled again. These people are so unreliable!

(5.19) [Different instances of a given type]
Every time I see you and Marshall together I think there's a happy marriage. I want one too. (COCA)

(5.20) [Bridging involving an event and its subevent]
For the second straight game, Ryan Hartman committed two penalties in the second period … (COCA)

(5.21) [Bridging involving an object and its part]
Who started the ridiculous big-grille trend, anyway? Audi, maybe? Not only is it an ugly look, but it's also exponentially absurd because most cars get their engine air through vents below the bumper, not through the grille. (COCA)

(5.22) [Bridging involving an event and one of its participants]
… We counted eight playoff series in which the home team won the first game but lost the last game. (COCA)

(5.23) [Bridging involving an event and one of its nonhuman props]
He remembers the electricity in the locker room before the game, the stadium's muddy field, the game's overall frustration … (COCA)

Challenge 8. *The sponsor for a RefEx can be vague*. For example, a speaker—let's call her Alice—might conclude an hour-long speech about the dangers of smoking by saying, "And *that's* why you shouldn't smoke!" Given that the whole speech was about smoking, when Alice gets to "that's why," which part of what she said is she actually referring to?

Clearly, she is including whatever reasons she gave not to smoke, and clearly, she is not including jokes or references to the temperature of the room. But if she presented vignettes about particular smokers' experiences, would those be among the reasons to avoid smoking? Could a typical listener even remember all the reasons or definitively tease apart the reasons from other speech content? Most likely not.

Challenge 9. *Even if a RefEx does not require a textual sponsor, it might have one—and the agent should track it down*. Some RefExes, such as well-known proper names and so-called *universally known* entities (e.g., *the universe*), do not require a textual sponsor. However, even if a RefEx does not require a textual sponsor, there are three good reasons for seeking one out anyway:

1. Universally known entities are lexically ambiguous. In addition to their universally known meanings—for example, *the atmosphere* referring to Earth's atmosphere—they have additional generic meanings. *The atmosphere* might refer to the atmosphere of some other planet or, metaphorically, to the collective mood of a group of people. In the best case, the meaning will be clear on first mention of the entity in a discourse, and all other mentions in the chain of coreference will take on the same interpretation.[12]
2. Proper names can have a large number of real-world referents. For example, *the CIA* can refer to the Central Intelligence Agency or the Culinary Institute of America, and the names *Henry*, *Dr. Adams*, and *Professor Tanenbaum* can refer to any number of real-world individuals. Ideally, the identity of any of these will be clear on first mention in a text, and all subsequent mentions will belong to the same chain of coreference.
3. Chains of coreference can help to compute discourse structure, which, as we said earlier, is a difficult and as-yet unsolved problem.[13] The idea is, if coreference links obtain between sentences, those sentences probably belong to the same discourse segment.[14] By contrast, if sentences do not contain any coreference links, there is likely to be a discourse boundary between them (i.e., the topic has shifted). Detecting discourse boundaries can be useful, for example, for dynamically establishing the window of coreference: one should look for sponsors only within the chunk of text that is about that particular thing.

Challenge 10. *RefExes and/or their sponsors can be elided*. Some types of ellipsis, like verb phrase ellipsis, are typical in any genre of English (5.24), while others, like subject ellipsis, are used only in stylistically marked contexts (5.25) or with the support of the nonlinguistic context (5.26).

(5.24) I like to wear running shoes if I can __. (COCA)

(5.25) She felt her heart beat. __ Felt the pulse of it against her face. (COCA)

(5.26) "It was an accident. I was aiming for that boy over there," she said and pointed to the ball field. (COCA)

Challenge 11. *There can be benign ambiguity with respect to the sponsor.* Benign ambiguity is ambiguity that does not impede reasoning. For example, in (5.27), the second *it* might be understood as the rose, the glass, the glass of water, or the glass of water containing the rose—all of which are functionally equivalent since, by the time we get to *it* in the sentence, the rose and the glass of water are already functioning as a unit.[15]

(5.27) She would also take her rose back to the office, put it in a glass of water, and place it on the windowsill. (COCA)

Challenge 12. *Anchoring RefExes in memory can be tricky.* As explained earlier, when a RefEx is encountered in a text, it might refer to something the agent already knows about (a known anchor), or it might refer to something new (and therefore require that a new anchor be created). The most obvious situation in which a RefEx should create a new anchor in memory is when an NP is introduced with an indefinite article (5.28) or certain quantifiers (5.29).

(5.28) Through the window, the more silent sounds of night. A dog barks somewhere. (COCA)

(5.29) Rachel was going over the company phone bill and saw calls to a number she didn't recognize, so she dialed it. Some girl answers. (COCA)

The LEIA knows that such expressions trigger the creation of a new anchor because the lexical senses for *a/an* and this meaning of *some* include a call to a procedural semantic routine called *block-coreference*. This instructs the agent to *not* seek a textual sponsor but, instead, to directly create a new anchor in memory during Situational Reasoning.

Note, however, that there are contexts in which an entity that is at first interpreted as new turns out to be already known, as in (5.30).

(5.30) I know someone was here and I know it was the same person who vandalized my car … (COCA)

We model coreference and memory management for such contexts analogously to a person's reasoning. When the LEIA encounters the string *someone*, it creates a new anchor in memory because this entity is being presented as new and unknown. The fact that it later becomes clear that it is actually known means that two different anchors need to be coreferred—or merged, depending on the memory management strategy. Thus, no special look-ahead strategy need be implemented to account for the possibility that an originally unrecognized entity is later recognized.

With many other types of referring expressions, it can be difficult to know when to link new information to existing memories. Say a LEIA encounters the following sentence:

(5.31) When Susan and Tom Jones met their 1930s Colonial, it wasn't exactly love at first sight. (COCA)

Although the name Tom Jones might immediately evoke the hero of a Henry Fielding novel—information that a well-read LEIA should know—this can't be the intended referent since the dates don't match up. And, in fact, the name Tom Jones is so common that a LEIA with even moderate real-world experience in text processing will likely have encountered multiple people named Tom Jones. For each person, the LEIA might know some distinguishing feature values—maybe age for one, profession for another, and the wife's name for a third—but it will certainly not know the full set of features that might be needed to decide whether incoming information relates to a known or new person with this name. In fact, even knowing that Tom Jones's wife is named Mary is of modest help since Mary, too, is a very common name, and there are many Tom and Mary Joneses in the world. In short, deciding which property values to take into account when determining whether a referent is known or new is a big challenge—one that is undertaken later, during Situational Reasoning.

Summary: To reiterate—and allow readers to test whether they can recall relevant examples—the dozen challenges presented above are as follows:

1. Detecting RefExes can be difficult.
2. The surface features of RefExes and their sponsors need not match.
3. Head-matching NPs do not necessarily corefer.
4. Some personal pronouns (e.g., *you*, *they*) can have a specific, generalized, or hybrid referent.
5. Fully interpreting RefExes can require making implicatures.
6. Coreferential categories need not be of the same syntactic category.
7. Referentially related entities need not be coreferential.
8. A RefEx's sponsor can be vague.
9. Even RefExes that don't require a sponsor might have one—and it should be sought out.
10. RefExes and/or their sponsors can be elided.
11. There can be benign ambiguity with respect to the sponsor of RefExes.
12. Anchoring RefExes in memory can be tricky.

5.1.4 Special Considerations about Ellipsis

Ellipsis is one method of realizing a referring expression. Its broadest definition is uncontroversial: it is the nonexpression of some meaning that can be recovered from the linguistic or real-world context. However, distinguishing ellipsis from related phenomena—such as fragments and telegraphic writing—is less clear-cut, as is classifying subtypes of ellipsis.

Our classification derives from a combination of linguistic principles, hypotheses about how people detect and resolve ellipsis, and the practical needs of building agent systems. We begin with some definitions.

Syntactic ellipsis is the grammatically licensed absence of an otherwise mandatory syntactic constituent. *Mandatory* is defined either by an argument-taking word's selectional constraints[16] (5.32) or by the grammar overall (5.33).

(5.32) [The multiword auxiliary *do not* requires a complement but permits its omission]
Both mammals (synapsids) and birds, reptiles and crocodiles (sauropsids) exhibit some form of mother-caring behavior, but amphibians do not __. (COCA)

(5.33) [The so-called *gapping* construction allows for ellipsis of the main verb in certain types of highly parallel coordinate structures]
Yet these forms invite the limpers to judge the runners; non-readers __, the readers; the inarticulate __, the articulate; and non-writers __, writers. (COCA)

Whereas syntactic ellipsis involves a constituent missing from a complete structure, a *fragment* is a subsentential constituent occurring in isolation. For example, in question-answer pairs like (5.34), the question sets up specific expectations about the answer, and the answer can be a fragmentary utterance that fills those expectations.

(5.34) "Where you going?" "North Dakota." (COCA)

Noncanonical syntax subsumes phenomena such as unfinished thoughts, interruptions, spurious repetitions, and self-corrections. It is not a type of ellipsis. Whereas both syntactic ellipsis and fragments are rule-abiding phenomena, noncanonical syntax is not. LEIAs treat it either by recovery procedures triggered by failures of syntactic analysis (during Pre-Semantic Integration) or by circumventing syntactic analysis and attempting meaning composition using an *ordered bag of concepts* methodology (during Situational Reasoning).

There are two subtasks in treating ellipsis: detection (identifying what's missing) and resolution (determining what it means). Both of these can be quite challenging, which explains why ellipsis has, to date, not received adequate attention in NLP systems, despite the extensive ink devoted to it in the theoretical literature.[17]

5.1.5 Wrapping Up the Introduction

If your impression at this juncture is, "There is more going on with reference than I had imagined," that is not surprising. One of the main challenges in treating the full spectrum of reference phenomena is organizing the treatment of all these phenomena across analysis modules in a way that (a) makes sense both theoretically and practically and (b) lends itself to enhancement in well-understood ways over time.

Readers have already seen the first stage of reference treatment: during Basic Semantic Analysis, all overt referring expressions, as well as many types of ellipsis, have been

detected and at least partially semantically analyzed. For example, *he* is understood as HUMAN (GENDER male), and an elided verb phrase (licensed by a modal or aspectual verb; e.g., *You didn't __?*) is detected and provisionally resolved as an underspecified EVENT. What remains is to ground these initial semantic interpretations in the discourse context. Often, this involves identifying their textual sponsors.

The rest of the chapter works through the textual-coreference procedures that the LEIA carries out at this stage, which rely on the linguistic, lexical, and ontological knowledge that it possesses for general domains. Later, during Situational Reasoning, the agent will have access to more types of heuristic evidence (e.g., the results of vision processing) and more types of reasoning (e.g., reasoning about its role in the task at hand). These will help it to finalize coreference decisions and anchor referring expressions in its memory.

As we said, there is no getting around the sheer number of phenomena to be treated, their complexity, or the reality that achieving high-confidence results for all of them will require a lot of time and work. Accordingly, our claims about the current state of this microtheory are relatively modest:

1. It organizes the treatment of a large number of phenomena in a way that is cognitively motivated and practically useful.
2. It significantly advances the treatment of several important phenomena, such as anaphoric (including elliptical) event coreference, broad referring expressions, and bridging constructions.
3. It integrates coreference resolution with lexical disambiguation. That is, once the agent knows, for example, what a pronoun refers to, it can use that information to help disambiguate the other words in its clause.

The microtheory has proven useful even in its current state. We estimate that dedicating one linguist-year to each of around twenty well-defined problems would advance this microtheory to a seriously operational level for English-speaking LEIAs.

Each section of the upcoming narrative addresses a different type of referring expression: personal pronouns (5.2), pronominal broad referring expressions (5.3), definite descriptions (5.4), anaphoric event coreference (5.5), other types of elided and underspecified events (5.6), and overt event references (5.7). The more detailed sections conclude with a recap that extracts the main points and examples. These can be skimmed as a memory refresher and road map.

5.2 Personal Pronouns

LEIAs use a three-stage process to analyze personal pronouns like *he* and *her*. First, they import the results of an externally developed coreference engine (5.2.1). Then they run an internally developed personal pronoun–coreference function, which offers higher-confidence resolution for certain types of examples (5.2.2). And finally, they evaluate all posited coreference links for their semantic suitability (5.2.3).

5.2.1 Resolving Personal Pronouns Using an Externally Developed Engine

When LEIAs use the Stanford CoreNLP tool set for preprocessing and syntactic analysis, they also run its coreference resolver (hereafter, CoreNLPCoref)[18] and store the results for use at this stage. CoreNLPCoref is comprised of precision-ordered sieves that identify sponsors for certain *instances* of certain *types* of referring expressions. It offers state-of-the-art results within the limitations of knowledge-lean methods. The details of CoreNLP-Coref are described in Lee at al. (2013).[19]

For many types of referring expressions, CoreNLPCoref treats only a subset of instances. For example, although it treats the pronoun *it*, it does not treat examples in which the sponsor is one or more clauses (5.35), and although it treats the pronoun *they*, it does not treat examples in which the sponsor must be composed from different constituents on the fly (5.36).[20]

(5.35) But I'm glad she's happy! It makes me happy![21] (COCA)

(5.36) A month or two later, she met another man. They fell in love and got married. (COCA)

There are also coreference phenomena that CoreNLPCoref treats infrequently if at all, such as ellipsis and demonstrative pronouns (*this*, *that*).

CoreNLPCoref does quite well at resolving reflexive pronouns, proper names, definite descriptions, and first- and second-person pronouns. However, its precision on third-person pronouns is much lower. This is not unexpected since the latter's interpretation often requires real-world knowledge and reasoning. In fact, the Winograd Schema Challenge (Levesque et al., 2012) posits that the interpretation of third-person pronouns in pairs of sentences like (5.37a,b) can be used to gauge the intelligence of artificial intelligence systems.

(5.37) a. Joan made sure to thank Susan for all the help she$_{\text{SUSAN}}$ had given.
b. Joan made sure to thank Susan for all the help she$_{\text{JOAN}}$ had received.

Since LEIAs need a confidence estimate for each language-processing decision, we convert the precision scores for each sieve reported in Lee et al. (2013) into estimates of confidence. For example, the precision of the Exact String Match sieve serves as an estimate of the LEIA's confidence in resolving each RefEx treated by that sieve.

5.2.2 Resolving Personal Pronouns Using Lexico-Syntactic Constructions

As just mentioned, the weakest aspect of CoreNLPCoref (apart from not treating some types and instances of referring expressions at all) is resolving third-person pronouns. Although one might think that the only next step is to invoke semantics and pragmatics, it turns out that we can squeeze a bit more predictive power out of surface (lexico-syntactic) features by bunching them into constructions. These constructions reflect the fact that language is quite formulaic, and that phenomena such as parallelism have strong effects throughout the language system.[22]

Below we describe six constructions that we developed and tested to assess the utility of knowledge engineering in this domain. It is important to note that the published results of this evaluation (McShane & Babkin, 2016a) addressed exclusively the more difficult personal pronouns (*he*, *him*, *she*, *her*, *they*, and *them*), not taking credit for the fact that the same constructions would also resolve the easier ones (*I*, *me*, *you*, *we*, *us*).[23]

In the descriptions of each construction, *feature matching* means having the same value for person, number, and gender. In all cases, the pronoun and its sponsor are at the same level of quotation—that is, they are both either within or outside a direct quotation. *Sequential* implies that there are no other categories of the given type intervening. Square brackets indicate the number of examples evaluated and the percentage correct in the evaluation.[24] Each construction is illustrated by an example from the English Gigaword corpus (Graff & Cieri, 2003).

Construction 1. Sequential string-matching subjects of coordinated clauses [20 examples; 100% correct].

(5.38) The Warwickshire all-rounder Roger Twose has been named in the New Zealand squad to tour India beginning in October. Now he has taken the decision to make his life in New Zealand and he goes with our blessing and best wishes. (Gigaword)

Construction 2. Sequential feature-matching subjects of speech act verbs [20 examples; 90% correct].

(5.39) Established Zulu actors were used in the dubbing process, which took more than a month, he said. He said senior politicians, among them PWV provincial premier Tokyo Sexwale, and leading movie personalities had been invited to the gala … (Gigaword)

Construction 3. Sequential string-matching subjects in a [main-clause + subordinate-clause] structure [32 examples; 100% correct].

(5.40) Rabin also accused Iran of controlling the Islamic fundamentalist group Hezbollah, which has been blamed for several terrorist attacks. But he said he believed the weapons flow through Syria had slowed in recent months. (Gigaword)

Construction 4. Sequential feature-matching subjects in a [main-clause + subordinate-clause] structure [20 examples; 85% correct].

(5.41) President Bill Clinton warned Saturday that he would veto any attempt by Republicans to scrap plans to put 100,000 additional police on US streets in line with his prized crime-fighting package. (Gigaword)

Construction 5. Sequential feature-matching subjects of identical verbs [20 examples; 90% correct].

(5.42) The survivors of the family live under one roof. They live frugally on rice and beans distributed by the church. (Gigaword)

Construction 6. Direct objects of sequential coordinate clauses [20 examples; 85% correct].

(5.43) In addition, some 170 US soldiers will go to Saudi Arabia to take two Patriot missile batteries out of storage and transfer them to Kuwait, the Pentagon said. (Gigaword)

To be clear, we are not saying that CoreNLPCoref would get these wrong. However, even if its answers were correct, it would have no way of knowing that these resolutions were more confident than the corpus-wide average for third-person pronouns. We are also not saying that the results of this small-scale evaluation are definitive: after all, one can come up with counterexamples for all these generalizations. What we *are* saying is that constructions are a useful tool for establishing higher-confidence resolutions for particular types of lexico-syntactic contexts.

Many of the mistakes detected during the abovementioned evaluation would be corrected by semantic analysis, which was not part of the evaluation setup reported in McShane and Babkin (2016a).[25] For example, in (5.44), Construction 4 incorrectly posited *tears of joy* as the sponsor for *they*:

(5.44) Tears of joy and grief poured from the two teams as they lined up for the medal ceremony. (Gigaword)

However, since tears of joy are inanimate and therefore cannot be the agent of lining up, semantic analysis (described in the next section) will reject *tears of joy* as the coreferent for *they* and opt for *the two teams* instead.

We expect that further knowledge engineering will yield more constructions that have better-than-baseline predictive power and, therefore, can better inform the LEIA's combined analysis of semantics and reference, to which we now turn.

5.2.3 Semantically Vetting Hypothesized Pronominal Coreferences

All of the coreference votes hypothesized so far have relied on surface (lexico-syntactic) features. At this point, the LEIA invokes semantics: it checks (a) whether the coreferences are semantically valid and (b) whether they can help to disambiguate other constituents in their clause. Consider example (5.45):

(5.45) Mike talked at length with the surgeon before he started the operation.

Anyone reading this sentence outside of context should conclude that *he* most likely refers to the surgeon who is about to start a medical procedure. This is not the coreference

decision that was offered by CoreNLPCoref or by our own constructions. Both of these corefer *he* with *Mike*. So, we want the LEIA—like a person—to override the surface-feature-based expectation using semantic reasoning. To understand how the LEIA does this, we must zoom out to the level of complete, multistage text analysis.

During Basic Semantic Analysis the LEIA computes four candidate interpretations of (5.45): either Mike or the surgeon could be starting either SURGERY or a MILITARY-OPERATION. The ontology indicates that SURGERY and MILITARY-OPERATION expect different kinds of AGENTS, as shown below:

SURGERY		
AGENT	default	SURGEON
	sem	MEDICAL-PERSONNEL
	relaxable-to	HUMAN
MILITARY-OPERATION		
AGENT	default	MILITARY-PERSONNEL
	sem	HUMAN

Based on this knowledge, the highest-scoring analysis of *he started the operation* will involve either a surgeon performing surgery or a member of the military performing a military operation, since these are the *default* AGENTS of their respective events. However, the given context makes no mention of a member of the military, so the only highest-scoring analysis that is available involves a surgeon performing surgery. There are two next-highest-scoring analyses: either Mike or the surgeon could be performing a military operation. These receive high (but not maximally high) scores because, as HUMANS, they fulfill the basic semantic constraint (listed on the *sem* facet) of the AGENT of MILITARY-OPERATION. The lowest-scoring option is that Mike is performing surgery, since it resorts to the *relaxable-to* facet for the AGENT slot: HUMANS can only in a pinch be the AGENTS of SURGERY (as when performing a lifesaving procedure on a battlefield). In short, the coreference relation that was preferred using surface heuristics was dispreferred using semantic ones, and semantic ones always win.

At least two questions might come to mind at this point. First, shouldn't reference resolution procedures for *the operation* have already determined which kind of operation it is? Ideally, yes, but (a) that decision, too, could be right, wrong, or as-yet undetermined (i.e., residually ambiguous), and (b) even without any previous context, people interpret the sentence as involving a surgeon performing surgery—and so, too, should LEIAs.

The second question is, What if something else entirely is going on, which might be indicated in the preceding context or might rely on information the interlocutor is expected to know? For example, Mike might be an anesthesiologist who starts the surgery by knocking out the patient before the surgeon cuts; or Mike might be a doctor who is planning to perform a small surgery but wants to check some details with a senior surgeon first; or Mike might be a general whose only trusted advisor happens to be the unit's surgeon, with

whom he consults before launching any military operation. Language and the world allow for all these interpretations—and so does our modeling strategy. However, such reasoning does not happen *at this point in the process*. At this stage, the LEIA is using general linguistic and ontological knowledge to semantically vet coreference votes and, ideally, come up with the same interpretation that a person would outside of context. Later on, during Situational Reasoning, the agent again vets the posited coreference votes using all its situational awareness and world knowledge. This models a person having access to the whole context and the entire shared knowledge space.

5.2.4 Recap of Resolving Personal Pronouns during Basic Coreference Resolution

1. Some instances of pronouns are resolved by the externally developed CoreNLPCoref system: for example, Jack asked Suzy to marry him.
2. Some instances of pronouns are resolved more confidently by our internally developed lexico-syntactic constructions: But he said he believed she was right.
3. All posited coreference links are semantically vetted and, if necessary, overridden: Mike talked at length with the surgeon before he started the operation.

5.3 Pronominal Broad Referring Expressions

Broad referring expressions are called *broad* because they can corefer with either a noun phrase or a larger span of text.[26] Broad RefExes are commonly realized by the pronouns *this*, *that*, and *it*, but certain full noun phrases—headed by words like *suggestion* and *proposal*—can have broad reference as well.[27] This section addresses only pronominal broad RefExes.

Practically no coreference systems attempt to treat broad RefExes.[28] Resolving them can be relatively simple, as in (5.46), or beyond the state of the art, as in (5.47). In fact, in (5.47) the referent for *this* is not even in the text; it must be understood as something like "this person's behavior and how it is affecting me."

(5.46) They don't trust us. That's good. (COCA)

(5.47) She picked up a fork, stared at the food for a moment, then shook her head in despair. Fear had taken away her appetite. This can't go on, she thought angrily. Whoever he is, I won't let him do this to me. (COCA)

Since LEIAs will not be able to resolve all instances of broad RefExes anytime soon, we are supplying them with a method to independently detect which ones they *can* resolve with reasonable confidence and resolve only those (recall the simpler-first modeling principles introduced in chapter 2). Any examples not treated at this stage will be reconsidered during Situational Reasoning, when more resources can be brought to bear.

The current model includes five methods for detecting and resolving treatable instances of pronominal broad RefExes. These methods are described in the subsections below. For evaluations and additional details, see McShane (2015) and McShane and Babkin (2016a).

5.3.1 Resolving Pronominal Broad RefExes Using Constructions

Section 5.2.2 showed how lexico-syntactic constructions can be used to resolve third-person pronouns. Such constructions are similarly useful for resolving pronominal broad RefExes. The constructions which we have tested to date are illustrated by the following set of examples, each one introduced by an informal presentation of the associated construction.[29]

(5.48) [Ask/Wonder why CLAUSE, it's/it is because …]
If you've wondered why so many 80- and 90-year-old women are named Alice, it's because the president's daughter was the inspiration for the most popular name for girls born in the early years of the century. (COCA)

(5.49) [Why AUX SMALL-CLAUSE? It's/it is because …]
"Why is he busy? It's because of the pressure that's being put on him," … (Gigaword)

(5.50) [If/When/In_each_case_where/Whenever/Anytime_when CLAUSE, it's/it is because …]
If Myanmar seems oddly quiet, it is because many are tired of struggle and just want to improve their lives. (Gigaword)

(5.51) [Not only AUX (NEG) this/it/NP_{subj} … , this/it_{subj} …]
"Not only did it show that the emperor was very much a human being, it was also a grim reminder of the defeat and subservience of their nation." (Gigaword)

(5.52) [This/It/NP_{subj} (AUX) not only …, this/it_{subj} …]
The module not only disables the starter, it shuts down the fuel injection so the car won't run … (COCA)

(5.53) [This/It/NP_{subj} has/had nothing to do with …. This/it_{subj} ….]
"This provision has nothing to do with welfare reform. It is simply a budget-saving measure," … (Gigaword)

(5.54) This/It/NP_{subj} is/was not about …. This/it_{subj} is about ….]
"This war is not about diplomacy," he added. "It is about gangsterism …" (Gigaword)

(5.55) [This is not (a/an/the) N …. It_{subj} is/it's …]
"This is not the lottery. This is this man's life, …" (Gigaword)

(5.56) [That's why … that's why …]
That's why we stayed in the game and that's why we won. (Gigaword)

These constructions reflect pragmatic generalizations:

- Asking *why* something happened is often followed by an indication of why: (5.48) and (5.49).
- *If ... then* constructions can be used to explain potential situations: (5.50).
- Saying that something *not only* does one thing often leads to saying what *else* it does: (5.51) and (5.52).
- Saying that something *does not* involve something often leads to saying what it *does* involve: (5.53)–(5.55).
- Repetition structures, which feature a high degree of parallelism, often contain coreferential expressions: (5.56).

Some broad RefExes have propositional sponsors, whereas others have NP sponsors. Although finding NP sponsors might seem simpler, it actually isn't since the LEIA does not know beforehand whether the sponsor it is seeking is an NP or a proposition.

Vetting these constructions on a corpus showed that some of them must be specified more precisely in order to exclude false positives.[30] For example, although (5.57) matches the construction described in (5.52), the sponsor for *it* is not *the Defender,* as would be predicted by that construction.

(5.57) And for the first time in history, the Defender not only wants to introduce its own new rule for the class of boat to be raced, but also to keep this rule secret. It will be disclosed to challengers at a much later stage, putting all challengers at a huge disadvantage. (Gigaword)

There are at least three reasons why this context should not be interpreted as matching the cited construction; each one involves disrupting the simplicity and parallelism that give the construction predictive power: the voice is different in the two parts—active in the first, passive in the second; there is a subordinate clause intervening between the two parts; and there are additional NPs in the first part that are also candidate sponsors. Clearly, for this and all of the other constructions, additional knowledge engineering is needed to determine rule-out conditions—that is, ways in which the context can become too complex for the constructions to predict coreference relations.

5.3.2 Resolving Pronominal Broad RefExes in Syntactically Simple Contexts

As we said, the sponsor for a broad RefEx can be an NP, a span of text representing one or more propositions, or a meaning that is not explicitly presented at all. In this section, we focus on detecting contexts in which (a) the broad RefEx refers to a span of text and (b) the boundaries of that span of text can be automatically determined with high confidence due to the syntactic simplicity of the context.

Consider examples (5.58) and (5.59), in which *that* can be resolved confidently based on the fact that the previous clause in each case is syntactically simple and its left-hand edge represents a natural boundary for the most local context—that is, it is a sentence break.

(5.58) They live far from their homes. That makes them stronger than if they formed a real community. (Gigaword)

(5.59) "Strong Serbia is not to the liking of some powers abroad, and that's why they are trying to break it up with the help of the domestic traitors," he said. (Gigaword)

By contrast, in (5.60) the sentence preceding *it* contains two conjuncts (i.e., two clauses joined by *and*). This offers two options as the sponsor of *it*: either the entire sentence or only the most recent clause.[31]

(5.60) "Police will go pass some prostitutes on the corner and harass some kids having a disagreement. It's because we're young." (COCA)

Often, as in this case, such contexts involve benign ambiguity: it doesn't make much difference to subsequent reasoning which resolution option is chosen.

In order to exploit the notion of *syntactic simplicity* in a computer system, we have to formally define it. We do so using an approach we developed when working on verb phrase ellipsis (McShane & Babkin, 2016b). The gist is that simple syntactic structures have only one main verb; this excludes sentences containing clausal conjunction, relative clauses, subordinate clauses, *if ... then* structures, and so on.[32]

In order to test the hypotheses that the system could (a) automatically identify syntactically simple contexts and (b) automatically predict the sponsor for broad RefExes in them, we focused on the following set of constructions, in which the italicized verbs could appear in any inflectional form:

- [simple clause] + despite this/that
- [simple clause] + because of this/that
- [simple clause] + this/that *is* why/because
- [simple clause] + this/that *means*, *leads to*, *causes*, *suggests*, *creates*, *makes*

The reason for targeting these constructions is that the broad RefExes in them tend to have a clause-level (rather than an NP) sponsor. In other words, this example extraction method returned a lot of examples relevant to the study and few false positives.

Our experimentation showed that this coreference resolution strategy worked well for the examples it treated, but it did not treat very many examples because of the strict constraints on what a *simple* clause could look like. In order to expand the system's coverage, we experimented with relaxing the definition of *simple* (see McShane & Babkin, 2016a, for details). The results were mixed. For example, if simple clauses are permitted to be

scoped over by modalities, then many more contexts are covered, but real-world reasoning is needed to determine if the modal meaning should be excluded from the resolution, as in (5.61a), or included in it, as in (5.61b).

(5.61) a. "I believe Jenny will swim faster than she ever has in Barcelona, and that means she has a good chance of bringing home five medals, though the color is still to be determined" … (COCA)

b. I believe Jenny will swim faster than she ever has in Barcelona, and that is why I bet big money on her.

Another way to expand the coverage of examples is to apply automatic sentence trimming to complex sentences. When trimming works correctly, it can turn nonsimple clauses into functionally simple ones. However, trimming is error-prone, often removing the wrong bits. We found this simplification method reliable only with respect to trimming away speaker attributions, as indicated by the strikethrough in (5.62).[33]

(5.62) "Energy efficiency is really the name of the game in terms of what we can do now," ~~she said, adding that she was disappointed that Bush did not adopt a more proactive stance on global warming, despite urging on the part of Blair.~~ "That's why today I'm calling on the president to show real leadership," she said, adding it was unacceptable to adopt a stance that other nations blamed for high greenhouse gas emissions, such as China and India, take steps first. (Gigaword)

We have just begun to investigate how to best relax the definition of *simple* in order to achieve higher recall without losing too much precision. This appears to be a promising area for additional knowledge engineering.

5.3.3 Resolving Pronominal Broad RefExes Indicating Things That Must Stop

People want bad things to stop. So, given utterances like *This must stop!*, *This is unacceptable!*, and *This is awful!*, one expects *this* to refer to something bad. The corresponding hypothesis we explored was as follows: If we compiled a list of *bad* events/states, and if the context immediately preceding a statement like *This is bad!* included something on that list, then that thing should be the sponsor for the broad RefEx.

Testing this hypothesis required a list of bad events, which in the current jargon are called *negative sentiment terms*. Although we found an automatically compiled list of this type (Liu et al., 2005), it included many words that were either not events or not necessarily negative, such as *gibe* and *flirt*. So, to support our experimentation, we manually compiled our own list of over 400 negative sentiment terms, using introspection combined with manual inspection of both WordNet and Liu et al.'s list.[34] Then we tested our hypothesis against this list using the constructions *It must stop* (5.63) and *This is unacceptable* (5.64), along with various synonymous and near-synonymous variations.[35]

(5.63) "This war is in no way acceptable to us. It must stop immediately" … (Gigaword)

(5.64) "… 1,200 people were detained and packed in here, in building 19–6. This is unacceptable in a member country of the Council of Europe" … (Gigaword)

We carried out this experiment as a stand-alone task, outside of the LEIA's semantic analysis process, so it oriented around word strings, not the ontological concepts used in TMRs.

Our corpus analysis suggested that these constructions are useful for predicting the sponsors for broad RefExes. Still, certain enhancements were needed to avoid false positives.

Enhancement 1. Negative sentiment terms should guide coreference resolution only if more confident strategies have failed. For example, lexico-syntactic parallelism has stronger predictive power than a negative sentiment term. So the sponsor for *it* in (5.65) should be identified using construction 5 from section 5.2.2, which predicts that NPs that are sequential subjects of identical verbs will be coreferential.

(5.65) "This incident is unacceptable to the national authority and to the Palestinian people and free world. It is unacceptable at all levels." (Gigaword)

Enhancement 2. Syntactic analysis is needed to avoid false positives, both when matching a construction and when matching the events in our list. For example, *It must stop to refuel* does not match our construction since the clause includes the complement *to refuel* (meaning that *it* must refer to a vehicle, not an event). Also, the system must identify the part of speech of a potential sponsor. Thus, *rebels* can serve as a sponsor for a broad RefEx only when used as a present-tense verb (*Every time I tell Charlie to be quiet in class he rebels. This must stop!*), not as a plural noun.

Enhancement 3. Broad RefExes may refer to multiple entities viewed as a set. Analysis of such contexts requires dynamic list concatenation. For example, in (5.66), all the underlined negative events must be concatenated into the sponsor for *it*.

(5.66) "The stories we are hearing of the harassment of political opponents, detentions without trial, torture and the denial of medical attention are reminiscent of our experiences at the hands of apartheid police. It must stop now" … (Gigaword)

In addition to investigating contexts that contained a readily identifiable bad event, we also investigated contexts that lacked such an event and found this generalization useful: *Any event that must be stopped must currently be going on*. There are at least three linguistic clues that an event is in progress: (a) a verb in the progressive aspect, (b) an adverbial expressing duration, and (c) a verb expressing an increase or a decrease in a property value (e.g., *grow ever louder*). For instance, in (5.67), the progressive aspect (*has been playing*) and the time adverbial (*for two hours straight*) suggest that *playing his recorder* is the sponsor of the broad RefEx, despite the fact that recorder playing can be quite nice if done well.

(5.67) That kid has been playing his recorder for two hours straight. This has to stop!

Of course, having multiple sources of evidence pointing to the same sponsor should increase the agent's confidence in its resolution decision.

The observation that bad things should stop is only one of many domain-independent generalizations that can guide the search for a broad RefEx's sponsor. Another obvious one involves the use of positive sentiment terms in the same way (*This must continue! This is fabulous/amazing!*). To reiterate a strategy from our discussion of the *theory-model-system* triad, we find it theoretically and methodologically preferable to use domain-independent generalizations and processing methods as much as possible. This strategy reduces LEIAs' cognitive load by allowing them to avoid using maximally deep and sophisticated reasoning for each problem they must solve.

5.3.4 Resolving Pronominal Broad RefExes Using the Meaning of Predicate Nominals

When a broad RefEx is used as the subject of a predicate nominal construction (i.e., *Broad-RefEx is/was NP*), it would seem that the meaning of the predicate nominal should indicate the semantic class of the sponsor. In some cases, this works well: in (5.68) *that* refers to a *year*, and 1971 is a year; in (5.69) *that* refers to a *place*, and *the prison* is a place.

(5.68) Back to 1971 for a moment. That was the year Texas Stadium opened, at a cost of $35 million. (COCA)

(5.69) The prison became for me the symbol of Soviet system. That was the place where there was an encounter between the last remnants of the freedom of Russia, between the last people who kept the survivors of freedom alive, and the leaders of the system which could be stable only if it controls the brains of all 200 million people. (COCA)

However, not all examples are so straightforward, as we found when we reviewed an inventory of automatically extracted examples of this kind. We identified the following five cases.

Case 1. When the predicate nominal is a proper name, there is almost never a textual sponsor: *This is World News Tonight with Peter Jennings*. (COCA)

Case 2. When the NP's meaning is vague, it does not usefully constrain the search for a sponsor. This was, unfortunately, the most common outcome in our investigation of the pattern *This/That [be] the [N]* using the online version of the COCA corpus. The most common vague (and, therefore, not useful) nouns were *way*, *problem*, *thing*, *reason*, *difference*, *case*, *point*, *subject*, *reason*, *reality*, *goal*, *theory*, *message*, *conclusion*. By contrast, the most common *useful* head nouns were *car*, *church*, *city*, *country*, *day*, *guy*, *location*, *man*, *person*, *place*, *plane*, *road*, *school*, *street*, *time*, *town*, *year*, *woman*.

Case 3. The optimal contexts contain exactly one candidate sponsor that either (a) matches the head noun of the NP, (b) is a synonym of that NP, or (c) is a hyponym or hypernym of that NP. For example, in (5.69), *prison* is a hyponym of *place*.

Case 4. In some cases, the context contains exactly one candidate sponsor whose identification should be possible given appropriate preprocessing. For example, proper name recognition is needed for *I love America*. *It is the place where I was born*; and date recognition is needed for example (5.68).

Case 5. Some contexts are tricky in some way—not necessarily too difficult to be automatically resolved using knowledge-based methods, but requiring some type of additional reasoning and/or resulting in some degree of uncertainty. Consider the examples below along with their analyses.

(5.70) [The sponsor is not grammatically identical to what is needed: *English* vs. *England*.]
But English soccer has a reputation it still can't shake off, no matter how hard it tries. This is the country that exported soccer violence back in the 1970s and '80s. (Gigaword)

(5.71) [The country *Czarist Russia* must be skipped over when working back through the text to find the sponsor.]
"My grandparents came to this country crammed into tight ship quarters from Czarist Russia because they believed this was the country where their votes would be counted" … (Gigaword)

(5.72) ['That' is whichever of the implied countries—the Czech Republic, Russia, Finland, or Sweden—has the most engaged big-time players.]
"Every year it is the same cast of characters, the Czechs, Russians, Finns and Swedes," Hitchcock said. "But it depends on the big-time players and if the big-time players are engaged then that is the country that wins." (Gigaword)

(5.73) [*The road* refers to a sequence of events that could end in another Great Depression.]
|Couric: If this doesn't pass, do you think there's a risk of another Great Depression? Palin: Unfortunately, that is the road that America may find itself on. (COCA)

An important detail is that, even if the system correctly points to the sponsor in these contexts, this does not fully resolve the meaning of the construction. For example, if *that* is resolved to 1971 in (5.58), the agent still has to combine the semantic interpretation of *1971 was the year* with the semantic interpretation of *Texas stadium opened*. We facilitate the agent's doing this by creating lexical senses for associated constructions. For example, the construction NP_{YEAR} *[be] the year (when/that)* EVENT will generate the meaning representation EVENT (TIME YEAR).

5.3.5 Resolving Pronominal Broad RefExes Using Selectional Constraints

Events require particular kinds of objects to fill their case roles. For example, the concept CELEBRATE is most commonly used with a HUMAN as the AGENT and a HOLIDAY as the THEME. This information is recorded in the agent's ontology as follows:

CELEBRATE		
AGENT	default	HUMAN
	sem	SOCIAL-ORGANIZATION
THEME	default	HOLIDAY

Moving from ontology to language, we expect the verb *celebrate* to be used in clauses like *They celebrated Thanksgiving together* and *Roberto celebrated his birthday*. Expectations like these can help to automatically resolve broad RefExes in some contexts. The study reported in McShane (2015) explored this potential in contexts like the following:

(5.74) Enron says the deal looks favorable because it was negotiated in 1992. (Gigaword)

(5.75) The water was drinkable because it boiled for several minutes. (Gigaword)

We hypothesized:

If a broadRefEx (e.g., *it*) fills a semantically highly constrained case role slot (e.g., the THEME of NEGOTIATE or BOIL)

And if a typical filler of that slot is available in the immediately preceding context (e.g., BUSINESS-DEAL or WATER),

Then that typical filler is the sponsor for the broadRefEx.

Our experimental setup did not invoke semantic analysis. It was carried out using a simpler methodology involving lists of verbs and keywords that typically filled their case role slots.[36] We found that, although the intuition was correct and useful, four enhancements were needed to improve the precision of this reference resolution strategy. We first provide a set of relevant corpus examples and then use them to illustrate the four enhancements.

(5.76) Residents said they were running out of food in a city that had its electricity cut two days ago. Some wounded Iraqis bled to death, and a family was buried under the ruins of their house after it was bombed by a U.S. jet, Saadi said. (Gigaword)

(5.77) A holiday honoring Vid, the ancient Slavic god of healing, has become one of the most fateful days on the Serb calendar. Now known as St. Vitus Day, it is celebrated June 28. (Gigaword)

(5.78) In the worst atrocity, some 5,000 men, women and children were slaughtered in the border town of Halabja in March 1988 when it was bombed and shelled with cyanide gas. (Gigaword)

(5.79) Although Imayev was talking about fighting and encirclements, Kuraly, a village of about 5,000 people, could not have appeared more peaceful. Like many Chechen villages, it was bombed by Russian airplanes during the fighting that started in December. (Gigaword)

(5.80) The 20-year-old started playing cricket at the Soweto Cricket Club soon after it was built 10 years ago. (Gigaword)

Enhancement 1. Syntactic analysis must be used to avoid false positive keyword analyses. For example, if keywords are part of a nominal compound, they must be the final (head) element of that compound. Thus, if a context contains "the restaurant garage.… It was bombed," then *garage*, not *restaurant*, is a candidate sponsor for *it.*

Enhancement 2. Candidate sponsors must be ranked according to recency, with the most recent being favored—even though recency is not a fully reliable heuristic. For example, although in (5.76) both a *city* and a *house* can be bombed, the sponsor for *it* is the more proximate *house*.

Enhancement 3. Chains of coreference must be identified: for example, in (5.77) *a holiday honoring Vid* and *St. Vitus Day* are in a coreference chain, so pointing to either of them is a correct resolution of *it*.

Enhancement 4. Certain preprocessing results must be included in the sponsor-selection heuristics: restrictive postmodification (5.78), appositives (5.79), and proper nouns with meaningful headwords, as in (5.80), where *Soweto Cricket Club* is an entity of the type SOCIAL-CLUB.

Outstanding problems involve the usual suspects, such as vagueness—in (5.81), is *it* the library or the palace?—and indirect referring expressions, such as metonymy (5.82).

(5.81) After the meeting, Kinkel and Mubarak inaugurated a public library in a renovated palace overlooking the Nile. It was built with a German grant of 5.5 million marks (dlrs 3.9 million). (Gigaword)

(5.82) The stolen van Gogh, he said, has special value because it was painted in the last six weeks of the artist's life. (Gigaword)

To sum up this subsection: When a verb's argument is realized as a pronoun, selectional constraints on this argument can guide the search for the pronoun's sponsor. There are multiple ways to operationalize this knowledge about the kinds of arguments required by different verbs. For the corpus analysis reported here, word lists were used. By contrast, during full semantic analysis by LEIAs, the combination of lexicon and ontology provides this knowledge.

5.3.6 Recap of Resolving Pronominal Broad RefExes

1. Pronominal broad RefExes occurring in listed constructions are resolved: "If you've wondered why so many 80- and 90-year-old women are named Alice, it's because …." (COCA)
2. Pronominal broad RefExes in syntactically simple (or pruned to become simple) contexts are resolved: "They live far from their homes. That makes them stronger than if they formed a real community." (Gigaword)
3. Pronominal broad RefExes referring to undesirable things are resolved: "'This war is in no way acceptable to us. It must stop immediately ….'" (Gigaword)

4. Pronominal broad RefExes described by predicate nominals are resolved: "Back to 1971 for a moment. That was the year Texas Stadium opened …." (COCA)
5. Pronominal broad RefExes filling narrow selectional constraints are resolved: "Enron says the deal looks favorable because it was negotiated in 1992." (COCA)

All the strategies used to resolve pronominal broad RefExes have three noteworthy features: (a) they do not require domain-specific knowledge or reasoning, so they can be applied to texts in the open domain; (b) they were developed and tested with only a small knowledge-engineering effort but still yielded quite useful results; and (c) they employ readily computable heuristics, which minimizes the cognitive load for LEIAs; this both simulates human functioning and promises to make LEIAs more efficient.

5.4 Definite Descriptions

Definite descriptions (NP-Defs, noun phrases with *the*) are treated at multiple stages of processing an input. Here we review what has already been done with them (section 5.4.1), present what is new at this stage (section 5.4.2), and describe what remains to be done at later stages (section 5.4.3).

5.4.1 Definite Description Processing So Far: A Refresher

The following steps occur during Pre-Semantic Analysis:

- CoreNLPCoref posts coreference votes for some instances of NP-Defs.
- The CoreNLP preprocessor identifies proper names with *the*, suggesting that they do not require a textual sponsor: for example, *the CIA*.

During Basic Semantic Analysis, these processes happen:

- The LEIA identifies nonreferring instances of NP-Defs, such as those used in idioms (*He kicked the bucket*). Since these words/phrases do not generate TMR frames, they are not subject to coreference procedures. If CoreNLPCoref has posited coreference votes for such words/phrases, they are ignored.
- For all TMR frames generated by an NP-Def, the LEIA creates a COREF slot filled with the call to the meaning procedure *resolve-NP-Def*. This happens because the meaning-procedures zone of the lexical sense of *the* contains this function call, which is copied into the basic TMR. For example, the TMR for the input *The horse is eating lazily* is as follows:

HORSE-1

AGENT-OF	INGEST-1
COREF	*resolve-NP-Def*
lex-sense	*horse-n1*
lex-sense-sub	*the-art1*

INGEST-1	
AGENT	HORSE-1
URGENCY	0
TIME	*find-anchor-time*
lex-sense	*eat-v1*
lex-sense-sub	*lazily-adv1*
lex-sense-sub	*be-aux1*

This TMR says that two lexical senses were invoked to create the HORSE-1 frame: *horse-n1* (the main one) and *the-art1* (a subservient one). It also indicates that two meaning procedures remain to be run—*find-anchor-time* (i.e., determine the time of speech, to account for the present tense) and *resolve-NP-Def* (i.e., find the sponsor for HORSE-1, since the presence of the article *the* indicates that it might need one).

In short, the basic TMRs for inputs that include *the* reflect three types of information: which instances of NP-Defs weren't referring expressions to begin with (there will be no TMR frames for them); which instances of NP-Defs are proper nouns that do not require, but might still have, a textual sponsor; and which instances of NP-Defs require additional reference-oriented processing. "Additional processing" does not mean that there is always a sponsor; instead, it means that either a sponsor must be found or the agent must understand why one is not needed.

5.4.2 Definite Description Processing at This Stage

This stage includes six functions for processing NP-Defs, described below.

5.4.2.1 Rejecting coreference links with property value conflicts If CoreNLPCoref identified a sponsor for the NP-Def, the LEIA checks whether that sponsor is plausible on the basis of any property values mentioned in the text. For example, *the blue car* and *the red car* cannot corefer, nor can *the Swedish diplomat* and *the Hungarian diplomat*. The reason these pairs cannot corefer is because the descriptions use the same property (COLOR or HAS-NATIONALITY) with different values (blue/red, Sweden/Hungary). By contrast, *the red car* can corefer with *the expensive car* because there are no conflicting property values: one description talks about COLOR whereas the other talks about COST.[37] That is, a red car can be expensive—no problem.

Note that, at this stage, only properties presented in the text itself are considered, not properties that are expected to be known as part of general world knowledge. For example, BMWs are expensive, high-quality cars, so it is unlikely that one would be referred to as *the cheap car* or *the poorly made car*. Later on, during Situational Reasoning, the agent will, yet again, check whether posited coreference links make sense, but at that point it will consult its knowledge of both the world and the situation.

5.4.2.2 Running reference-resolution meaning procedures listed in lexical senses Certain definite descriptions require special reasoning to be resolved: for example, the sponsor

for *the couple* must be two individuals, and the sponsor for *the trio* must be three. A special resolution function is recorded in the meaning-procedures zone of each such multiword lexical sense. The calls to these procedures are copied into the TMR during Basic Semantic Analysis and are run at this time.

5.4.2.3 Establishing that a sponsor is not needed As described in the introduction to this chapter, in some cases, NP-Defs do not require a sponsor. These cases include the following:

- *Universally known NP-Defs*: This covers such entities as *the earth*, *the sun*, and *the solar system*. We keep an ever growing list of such entities.[38] In the lexicon, each such noun includes at least two senses: one that requires *the* and refers to the universally known meaning, and another that does not require *the* and refers to a more general, related meaning.[39] Both of these analyses are generated during Basic Semantic Analysis. If CoreNLPCoref has suggested a sponsor for the given NP-Def, and if that sponsor has not been invalidated based on property values, then this analysis takes precedence over the candidate analysis "universally known and without a sponsor." Of course, a chain of coreference might employ the *universally known* interpretation—but that will be established by the *first* instance of NP-Def in a text, not the subsequent ones.
- *NP-Defs with restrictive modification*: Restrictive modifiers provide essential, non-optional, information about a noun.[40] These modifiers make the meaning of the noun phrase concrete in the real world, so there is no need for a sponsor. Restrictive modifiers can be detected using templates (stored as multiword senses of *the*), such as the following:
  - the + N + PP: *the streets of my hometown*
  - the + proper-N + N: *the French army*
  - the NUMBER + TEMPORAL-UNIT + (since/that/when) + CLAUSE: *Not exactly statuesque at five feet, but she'd grown a good three inches in the two years or so since he'd saved her from slavery.* (COCA)
- *NP-Defs that are proper nouns*: These will already have been identified by the CoreNLP named-entity recognizer, but it is at this point that the agent fills the COREF slot of the TMR with the filler *no-sponsor-needed*. (Recall that, although CoreNLPCoref is run during Pre-Semantic Analysis, the agent only consults its results at this stage.)
- *Generic uses of NP-Defs*: These are used primarily in definitional statements such as Wikipedia's *The lion (Panthera leo) is a species in the family Felidae*.[41]

5.4.2.4 Identifying bridging references Bridging references connect two entities that are not *co*referential but are semantically related in particular ways. Mentioning the

sponsor in the context virtually introduces semantically related entities, and they can be referred to using *the*. At this time, our model of bridging references covers bridging via object meronymy and event scripts.

Bridging via object meronymy. Meronymy is the has-as-part relation. Object meronymy means that one object is part of another object. Since we all know, for example, that windows can be parts of offices, we can say things like *I walked into her office and the window was open*. We use *the* with *window*—even though no window was mentioned before—because the potential for a window was effectively introduced into the discourse when *office* was mentioned. In the ontology, object meronymic relations are recorded using the HAS-OBJECT-AS-PART relation, for example,

OFFICE-ROOM
  HAS-OBJECT-AS-PART WINDOW, DESK, CEILING, DOOR, …

To detect object meronymy, the agent checks whether the ontological concept for the RefEx (here, WINDOW) is listed in the HAS-OBJECT-AS-PART slot of any of the ontological concepts used within the window of coreference (i.e., the immediately preceding context). The answer is yes for OFFICE-ROOM. As long as the ontology lists the needed meronymic relationship, this analysis is straightforward.

When object meronymy explains a given use of *the*, two things happen to the nascent TMR:

- The RefEx's frame is supplemented by the meronymic information—here, WINDOW-1 (PART-OF-OBJECT OFFICE-ROOM-1).[42]
- The COREF slot in the RefEx's frame, whose appearance was originally triggered by the lexical sense of *the*, is removed, showing that there is no coreference and none is needed.

Bridging via an event script. Scripts are complex events—that is, they are events that contain subevents. For example, AIR-TRAVEL-EVENT includes subevents such as a MOTION-EVENT that takes the flier to the airport, followed by AIRPORT-CHECK-IN, AIRPORT-SECURITY-CHECK, BOARD-AIRPLANE, AIRPLANE-TAKE-OFF, and so on.[43] The main event and all of its subevents have expected participants and props—in our example, PILOT, FLIGHT-CREW, AIRPLANE, AIRPLANE-TRAY-TABLE, and so on. The mention of these events virtually introduces all relevant participants into the discourse—and, therefore, they can be referred to using an NP-Def. For example, all the underlined instances of NP-Def in (5.83) are licensed by the mention of *flight*:

(5.83) I had an awful flight last week. It was bumpy and the pilot didn't explain why. The tray table was broken and kept falling on my knees. The flight attendant was in a bad mood and was snarky with everyone. And the landing gave me whiplash.

When script-based bridging explains a given use of *the*, it can lead to modifications of the nascent TMR, depending on the ontological relationship between the sponsor and the

RefEx. If they are directly linked in the script by a relation, then that relation is used. When *flight* (which is described in its TMR as FLY-PLANE-1) licenses *the pilot*, the frame for *pilot* is expanded to

```
PILOT-1
    AGENT-OF          FLY-PLANE-1
```

If, by contrast, the two entities are linked by a longer ontological path, then we use the generic RELATION instead. For example, we said above that *flight* can license *the tray table*. However, the actual ontological path between them contains multiple steps:

```
AIRPLANE-TRAY-TABLE-1
    LOCATION          AIRPLANE-1

AIRPLANE-1
    INSTRUMENT-OF     FLY-PLANE-1
```

The reason for this simplification is that our main goal at this point is to figure out why *the* was used. We can attain that goal, and record the results, without the additional complication of computing an ontological path. Impressionistically, we as humans understand that a tray table is related to a flight; it is not necessary to flesh out *how* when computing the meaning of our example text. As with the case of object meronymy, when we have explained the use of *the* by script-based reasoning, we remove the COREF slot because we are no longer looking for a coreferential sponsor.

5.4.2.5 Creating sets as sponsors for plural definite descriptions Sometimes the sponsor for plural definite descriptions must be dynamically composed from constituents, usually NPs, located in different parts of the text. In some cases, this is not difficult, as when the preceding context contains exactly two entities of the needed semantic type—in (5.84), NATION, and in (5.85), YEAR.

(5.84) For instance, in a well-known 1985 incident, a Coast Guard icebreaker navigated through the Northwest Passage, which the United States claims is an international strait, without seeking Canadian permission. In response, Canada "granted permission" (despite the lack of a request to that effect) for the voyage and, although the two countries agreed to the presence of Canadian observers onboard, the United States still disputed the Canadian claim of sovereignty over the waters. (COCA)

(5.85) Compared to the 1995 season, the 1996 season was strikingly different. The fleet was about 10% smaller than in 1995 and spent 23% fewer days fishing. Fishing was concentrated in the eastern region, whereas in 1995 it was concentrated in the western region (72%) between Papua New Guinea and Federated States of Micronesia (Fig. 3). Statistics on the number of sets/trip and trips/vessel, on the other hand, were essentially identical for the two years (Table 1). (COCA)

However, many cases are more difficult. For example, a context that refers to many people involved in a lawsuit—lawyers, a judge, defendants, spectators, witnesses—can conclude with, *In the end, the men were acquitted*. Interpreting *the men* requires reasoning that is more sophisticated than creating a set of all the HUMANs mentioned in the preceding context.

Our current algorithm for creating sets is rather simple:

- Identify the ontological class (i.e., concept mapping) of the NP-Def.
- Find all instances of that type, or its ontologically close hyponyms or hypernyms, in the window of coreference.
- If there are exactly two matching instances, create a set from them and corefer that set with the NP-Def.
- If there are not exactly two matching instances, postpone this resolution until Situational Reasoning.

5.4.2.6 Identifying sponsors that are hypernyms or hyponyms of definite descriptions We can refer to entities in more generic or more specific ways. The more generic references are *hypernyms* of the more specific ones; the more specific references are *hyponyms* of the more generic ones. For example, in a given context, *lung cancer*, *the cancer*, *the disease*, and *the affliction* can all refer to the same thing—for example, (5.86)—just as *the studio apartment*, *the apartment*, and *the residence* can all refer to the same thing.

(5.86) More than 30 years after declaring war on cancer, the disease refuses to surrender. (COCA)

The CoreNLPCoref engine that was run earlier does not identify sponsors for most contexts that require reasoning about subclasses and superclasses. LEIAs, by contrast, attempt such reasoning. We will illustrate how using example (5.86).

During Basic Semantic Analysis the LEIA will have generated candidate TMRs that analyze *cancer* as both CANCER-DISEASE and CANCER-ZODIAC. The only analysis for *disease* will be DISEASE (our current lexicon has only one non-construction-oriented sense of this word). The LEIA will check the ontology to see if either the pair [CANCER-DISEASE, DISEASE] or the pair [CANCER-ZODIAC, DISEASE] are in the same line of inheritance (i.e., if they are related as hypernym and hyponym). Only the former pairing is in the same line of inheritance, which both explains the use of the NP-Def *the disease* and allows the LEIA to disambiguate the analysis of *cancer*—all at one time.

The more specific the entities are (i.e., the deeper they are in the ontology), the more reliable the coreference-oriented reasoning. CANCER-DISEASE and DISEASE are quite specific, so there is a good chance they are coreferential. By contrast, if one of the entities is referred to as *the thing*—which maps to OBJECT in the ontology—it could corefer with any object

in the text. Of course, the agent still must resolve inputs that include *the thing* along with many other vague referring expressions. However, this cannot be done reliably without invoking more knowledge, which, for our agents, is done during Situational Reasoning.

5.4.3 Definite Description Processing Awaiting Situational Reasoning

By the end of this stage of processing, many instances of NP-Defs will have been resolved. Processing the remaining ones requires additional types of knowledge and reasoning that will become available during Situational Reasoning. These include the following:

- Interacting with the visual/physical environment. For example, if someone tells a robot, *Give me the hammer*, the robot needs to identify the intended hammer in the physical environment.
- Using features from long-term memory that are missing in the language context. For example, a speaker might refer to *the sisters* under the assumption that the listeners know which individuals in the context are sisters.
- Mindreading—that is, making inferences about the speaker's plans and goals, about shared knowledge, and so on. For example, if someone comes into work and the first thing he says is *The box finally arrived*, this assumes that both of the participants are aware that a box was expected.

5.4.4 Recap of Definite Description Processing at This Stage

The following types of NP-Def processing are carried out at this stage:

1. Reject coreference links with property value mismatches: for example, *the red car* and *the blue car* do not corefer.
2. Run the reference resolution functions recorded in the lexical senses for specific words and phrases: for example, *the couple*, *the trio.*
3. Establish when a sponsor is not needed—for example, for universally known entities (*the sun*) and NP-Defs with restrictive postmodification (*the streets of my hometown*).
4. Identify bridging sponsors. For example, in the sentence *I walked into her office and the window was open*, *office* is the bridging sponsor for *the window.* When this is recognized, the TMR frame for WINDOW-1 will be enhanced to include PART-OF-OBJECT OFFICE-ROOM-1.
5. Create a set as the sponsor for plural NP-Defs: *Violists can be jealous of violinists for getting better parts, but the musicians still have to cooperate*.
6. Identify sponsors that are hypernyms or hyponyms of an NP-Def: *Go to the bank, the building on the corner.*

5.5 Anaphoric Event Coreference

Anaphoric event coreferences (AECs) can be expressed using verb phrase ellipsis[44] (5.87) or overt anaphors, such as *do it/that/this/so* (5.88).[45]

(5.87) Four family members can vote, and they all intend to __. (COCA)

(5.88) We want to keep the plant open—but we can't do it by ourselves. (COCA)

Both elided and overt-anaphoric verb phrases are detected during Basic Semantic Analysis using special lexical senses. Elided VPs are detected using lexical senses of modal and auxiliary verbs that anticipate a missing complement. Overt-anaphoric VPs are detected using the multiword senses for *do it*, *do that*, *do this*, and *do so*. The sem-strucs of those senses generate an underspecified EVENT in the TMR, which is flagged as requiring coreference resolution. This is, by the way, the same processing flow as is used for personal pronouns like *he* and *she*.

Fully resolving AECs involves answering up to five semantic questions, which we illustrate using the VP ellipsis–containing sentence in (5.89).

(5.89) John washed his car yesterday but Jane didn't __.

1. What is the verbal/EVENT head of the sponsor? Wash/WASH.
2. Do the elided event and its sponsor have instance or type coreference? Type coreference: there are two different instances of the WASH event.
3. Do the internal arguments have instance or type coreference (i.e., the same or different real-world referents)?[46] Type coreference: there are two different instances of AUTOMOBILE.
4. Are the meanings of modifiers in the sponsor clause copied or not copied into the resolution? Copied: *yesterday* applies to both propositions.
5. Are modal meanings in the sponsor clause copied or not copied into the resolution? This example does not contain modal meanings, but a slight variation on it does: *John tried to wash his car yesterday but Jane didn't.* In this case, the modal meaning *tried to* would be copied into the resolution: *Jane didn't try to wash her car yesterday.*

The results of this reasoning are shown in table 5.2, where the TMR frames are arranged to highlight the parallelism across the clauses. Note specifically that even though the second clause does not include the strings *wash*, *her car*, or *yesterday*, all these meanings are reconstructed during ellipsis resolution. So the meaning representation reads as if the input were "John washed his car yesterday but Jane didn't *wash her car yesterday*."

The subsections below further explore questions 1–5 above.

Table 5.2
Ellipsis-resolved meaning representation for *John washed his car yesterday but Jane didn't*

John washed his car yesterday but			Jane didn't*		
WASH-1			*WASH-2*		
AGENT	HUMAN-1		*AGENT*	*HUMAN-2*	
THEME	AUTOMOBILE-1		*THEME*	*AUTOMOBILE-2*	
TIME	(*combine-time* -1 day)		*TIME*	*(combine-time -1 day)*	
			SCOPE-OF	*MODALITY-1*	
HUMAN-1			HUMAN-2		
HAS-PERSONAL-NAME		'John'	HAS-PERSONAL-NAME		'Jane'
GENDER		male	GENDER		female
AGENT-OF		WASH-1	AGENT-OF		*WASH-2*
AUTOMOBILE-1			*AUTOMOBILE-2*		
RELATION	HUMAN-1	; "his"	*RELATION*	*HUMAN-2*	
THEME-OF	WASH-1		*THEME-OF*	*WASH-2*	
CONTRAST-1		; "but"	MODALITY-1		
DOMAIN	WASH-1		TYPE	epistemic	
RANGE	WASH-2		VALUE	0	; negation
			SCOPE	*WASH-2*	

* The elements in italics are reconstructed through VP ellipsis resolution.

5.5.1 What Is the Verbal/EVENT Head of the Sponsor?

Identifying the verbal/EVENT head of the sponsor is the starting point for all the other aspects of AEC resolution. Identification can range from very simple to very difficult, as illustrated by the contrast between (5.90) and (5.91). Note that in this section, since we are focusing on identifying the head of the sponsor clause, we underline only the head in the examples.

(5.90) I'm 51 now, and if I'm going to do it I'd better do it now. (COCA)

(5.91) "You have to guard students and weed out the bad kids and get them in an alternative setting or get out of school entirely," Roach said. "And there are federal laws with special education that prevent you from doing that." (COCA)

Finding the sponsor head for the second *do it* in (5.90) is easy because the context has the following combination of coreference-predicting property values (recall section 2.6 on simpler-first modeling):

- The coreferential RefExes are identical (*do it ... do it*), which reflects parallelism and simplicity.

- The coreferential RefExes are in an *if ... then* construction, which reflects prefabrication.
- The two parts of the construction contain one simple proposition each, which reflects simplicity.
- This expression, or close paraphrases of it (*If* someone$_i$ *is going to do it,* pronoun$_i$ *(had) better do it now/soon/fast*), should be familiar to every native speaker of English, which reflects prefabrication and ontological typicality.

Note that the first instance of *do it*, which serves as the sponsor for the latter instance, must itself be resolved as well, but that is a separate task that might be simple or complex.

By contrast, (5.91) is complex—one needs to apply human-level reasoning to understand what is disallowed. We, in fact, are not sure what the intended resolution is. It could be any of the propositions labeled A, B, C, and D in (5.92).

(5.92) You have to [$_A$ guard students and [$_B$ weed out the bad kids and [$_C$ get them in an alternative setting or [$_D$ get out of school entirely]]]].

As with all of our microtheories, we first focused on identifying the subset of cases that could be treated using heuristics that could be applied to any text, without requiring that the agent have specialized domain knowledge (like the federal laws about special education) or reasoning capabilities. These easier cases are automatically detected by the agent and resolved at this stage. All residual examples are postponed until Situational Reasoning.

We implemented two different versions of the sponsor-head-identification algorithm (reported in McShane & Babkin, 2016b; McShane & Beale, 2020) that relied on similar theoretical principles—simplicity, parallelism, and prefabrication. They were operationalized into quite different models and associated systems, the latter having a more developed theoretical substrate. The basic insight of both is that in some constructions, syntax alone can predict the sponsor for AECs. This harks back to the cornerstone observation of generative grammar: that grammaticality judgments can be made even for nonsensical sentences. So, even for a gibberish sentence like (5.93), we can easily reconstruct the elliptical gap as *gwaffed the gappulon*.

(5.93) The sloip gwaffed the gappulon and the loips did __ too.

The question is, how far can we push syntax-only sponsor-head selection before needing to resort to semantics? Or, stated differently, how can we formally capture the perceived simplicity of examples like (5.94)—(5.97), so that agents can confidently identify their sponsor heads, while postponing examples that require semantic reasoning?

(5.94) I want to do it and it would make me sad if I didn't __. (COCA)

(5.95) We needed to match Soviet technology for national defense purposes, and most Americans understood the dangerous consequences if we did not __. (COCA)

(5.96) In my opinion, none of these quarterbacks coming out should play in 1999, but you know some of them will __. (COCA)

(5.97) It's your money; do whatever you want to with it. If you want to make a big pile of it and burn it, you can do it. (COCA)

Note that although these examples are relatively simple—and were correctly treated by the model/system described in McShane and Beale (2020)—they are not trivially simple. For example, in resolving (5.94)–(5.96), the system needed to ignore the most proximate verbal candidates (*make/understood/know*) as well as exclude the sponsor-clause modals (*want/needed to/should*) from the resolution. And in resolving (5.97), the system needed to include a conjoined VP in the resolution, while excluding the sponsor-clause modal.

The goal of this research was to see how far we could push a lexico-syntactic approach before it broke, since focusing on only the absolutely simplest cases would offer too little coverage of examples to be very useful. Although both our 2016 and 2020 models focused on identifying the simpler cases, the models themselves are not simple. That is because not only did all of the component heuristics need to be formally defined in terms of the processors that could be used to compute them, but the limitations of those processors affected the extent to which the models could, at the current state of the art, be faithfully captured in an implementation.

Of course, syntax is not the only tool LEIAs have available at this stage for AEC coreference resolution. They can also apply selectional constraints unilaterally to attempt to understand what the underspecified EVENT could, in principle, mean and, therefore, which candidate sponsor is most fitting. For example, the sponsor head for the elided VP in (5.98) could be either *reading* or *munching*. But since dogs can't read, *munching* is the clear-cut semantically informed choice.

(5.98) Carol was reading a book and munching on corn chips. Her dog was __ too.

5.5.2 Is There Instance or Type Coreference between the Events?

If the elided event precisely corefers with its sponsor, there is *instance coreference* (5.99a), whereas if the reference is to a different *instance* of the same *type* of event, there is *type coreference* (5.99b).

(5.99) a. Jim tried to open the bottle but couldn't __.
b. Jim couldn't open the bottle but Jerry could __.

Note that even though the same bottle is being opened, as long as there are different agents, there must be different event instances. That is, event-instance coreference requires that

all property values unify. However, either of the clauses can include additional modifiers, as in (5.100).

(5.100) Jim tried to open the bottle but couldn't __ without a bottle opener.

Our current algorithm for making the type-versus-instance coreference decision is as follows:

If the clauses have non-coreferential subjects; *This was determined earlier*
Then there is type coreference
Else
  If the clauses have coreferential subjects
  Then check all other property values
    If none have conflicting values
    Then this is instance coreference
    Else (i.e., if some property values conflict) it is type coreference.

5.5.3 Is There Instance or Type Coreference between Objects in the VPs?

If the sponsor clause includes objects (a direct object, an indirect object, or the object of a preposition), their reconstructions in the AEC clause might precisely corefer with those in the sponsor clause, which is instance coreference (5.101), or there might be different instances of the same type of object, which is type coreference (5.102). Starting with this set of examples, the entire sponsor is underlined.

(5.101) [Same fence, same alley]
I jumped the fence into the alley and Sally did __ too.

(5.102) [Sally walks her dog]
I walk my dog every morning and Sally does __ too.

Determining the instance-versus-type coreference of internal arguments (direct and indirect objects) is tricky and can require reasoning about how things generally work in the world. Our current algorithm is as follows:

Use instance coreference for internal arguments if either of these is true:

- The event coreference is instance coreference: John made this pizza, he really did __.
- The event coreference is type coreference and the sponsor-clause internal argument is NP-Def: He jumped over the fence and I did __ too.

Use type coreference for internal arguments if any of these are true:

- The event coreference is type coreference and the sponsor-clause internal argument has *a/an/some*: He ate a sandwich and I did __ too.

- The event coreference is type coreference and the sponsor-clause internal argument has a possessive modifier (*his*, *her*, *Martin's*, and so on): John washed his car and I did __ too.
- The event coreference is type coreference and the sponsor-clause internal argument has no article or determiner: John watches birds and I do __ too.

5.5.4 Should Adjuncts in the Sponsor Clause Be Included in, or Excluded from, the Resolution?

If the AEC clause specifies a value for PROPERTY-X, and if the sponsor clause *also* has a value for PROPERTY-X, then the sponsor clause's value is not copied during AEC resolution. For example, in (5.103), the interpretation of *today* fills the TIME slot in the AEC clause TMR, blocking the copying of the interpretation of *yesterday*.

(5.103) He arrived by train yesterday and she did __ today.

All other property values—here, *by train*—are copied over during AEC resolution, and, if applicable, the instance-versus-type coreference rules described above are invoked. For example, in (5.104), the meaning of *by the pool* fills the LOCATION slot and uses the same instance of SWIMMING-POOL because *the pool* is NP-Def.

(5.104) Grandma was drinking wine by the pool and Grandpa was __ too.

5.5.5 Should Modal and Other Scopers Be Included in, or Excluded from, the Resolution?

The algorithm for identifying AEC sponsor heads relies heavily on modality as a heuristic (see McShane & Beale, 2020, for details). For example, the pairing of the modals *might* and *might not* in (5.105) suggests not only that the sponsor head is *help* but also that the modal *might* should be excluded from the resolution (we don't want to end up with *But it also might not might help the cod*).

(5.105) It might help the cod. But it also might not __. (COCA)

Apart from modalities, other meanings can scope over propositions and be either included in, or excluded from, AEC resolutions:

(5.106) a. Whereas he vowed to tell the truth, she actually did it.
b. Whereas he vowed to tell the truth, she didn't __.

(5.107) a. He said he would come but he didn't __.
b. He said he would come but she didn't __.

The best way to study how scopers work is to collect corpus examples of all the kinds and combinations of scopers in the sponsor clause and the ellipsis clause, respectively. As a first

step toward this, we created a list of nearly one hundred modal/aspectual correlations and searched for examples of them in the Gigaword corpus (Graff & Cieri, 2003). The search patterns grouped together inflectional forms of words and synonyms such as the following:

doesn't V … doesn't try to __

can start $V_{PROGRESSIVE}$ … won't __

As it turned out, many of the correlations were not attested at all. Others had many hits, but their behavior was entirely predictable a priori, leading to no new insights. For example, given a positive-negative pair of modal verbs (e.g., *could/couldn't*), the sponsor-clause modal should be excluded from the resolution. But some of the study results did suggest the need for specific rules, as illustrated by the following examples:

(5.108) a. They didn't want to go so they didn't __.
b. They didn't have to go so they didn't __.

(5.109) a. They didn't want to go and we didn't __ either.
b. They didn't have to go and we didn't __ either.
c. They didn't try to go and we didn't __ either.

In all these examples, the modal verb *didn't* occurs in both the sponsor clause and the ellipsis clause, and *didn't* in the first clause scopes over another modal (*want to*, *have to*, *or try to*) as well as the head verb (*go*). The key difference between the two sets of examples is that in (5.108), the clauses' subjects are coreferential, whereas in (5.109), they are not. When the subjects are coreferential, the additional modal (*want to*, *have to*) is excluded from the ellipsis resolution; by contrast, when the subjects are different, the additional modal is included in the resolution. From the human perspective, we can say that these resolution choices simply make sense. But from the system's point of view, these decisions need to be recorded as resolution rules.

At the time of writing, a more comprehensive microtheory of scoper treatment is under development. It needs to address not only modal and aspectual verbs but also any other words and multiword expressions that can take a verbal complement.

5.5.6 Recap of Anaphoric Event Coreference

To identify the sponsor for an elided or anaphoric event, the following questions must be answered:

1. What is the verbal/EVENT head of the sponsor?
2. Is there instance or type coreference between the events?
3. Is there instance or type coreference between objects in the VPs?
4. Should adjuncts in the sponsor clause be included in, or excluded from, the resolution?
5. Should modal and other scopers be included in, or excluded from, the resolution?

5.6 Other Elided and Underspecified Events

During Basic Semantic Analysis (the previous stage of processing), the agent (a) detected various types of ellipsis, (b) provisionally resolved them using generic EVENTS as placeholders, and (c) put flags in the TMRs indicating that their meanings needed to be specified. Those flags are, in fact, calls to procedural semantic routines that require various types of heuristic evidence that become available at different times. Here we consider two types of verbal ellipsis whose meaning can often be computed at this stage.

Aspectuals + OBJECTS. When an aspectual verb (e.g., *start*, *finish*, *continue*) takes an OBJECT as its complement (e.g., *He started a book*), this is a clear sign that an EVENT has been elided: after all, one can only start, finish, or continue *doing something* with an object.

One method of establishing the default interpretation in such contexts is querying the ontology. The question for the agent is, Are there any EVENTs for which the default meaning of the subject is the AGENT and the default meaning of the object is the THEME? Recall that, in the LEIA's ontology, case role fillers are described using facets that represent three levels of constraints: the default constraint (*default*), the basic semantic constraint (*sem*), and the expected potential for extended usages (*relaxable-to*). If there is exactly one EVENT whose *default* case role fillers fulfill the search criteria, then that event is a strong candidate as a default interpretation. This would occur if, during the analysis of the sentences *She started a book* and *The author started a book*, the LEIA found the following in the ontology:

```
READ
    AGENT       default    HUMAN
    THEME       default    BOOK-DOCUMENT

WRITE
    AGENT       default    AUTHOR
    THEME       default    BOOK-DOCUMENT
```

However, what if the ontology contained different information? What if it listed HUMAN as the default agent of WRITE? In that case, both sentences would be analyzed, by default, as referring to either reading or writing.

Apart from the complexities of decision-making during ontology acquisition (there often is no single perfect answer), many other details need to be worked through to achieve a strongly predictive microtheory of *aspect* + OBJECT resolution. For example, the sentence itself might not present the most specific possible information. The *she* in *She started a book* might refer to a well-known author, and her social role may or may not be relevant in the given context (she might be starting to read a book on an airplane, just like a nonauthor would do). The full inventory of contextually relevant properties will become available during the later stage of Situational Reasoning.

Conditions of change. Events are caused by events. If a text says that an *object* causes an event, this is a sure sign that a related event has been elided. For example, *The onions*

made her cry actually means something like *Molecules released when the onion was cut came in contact with the tissues of her eye, causing a chain of events that led to crying.* Or, for someone who hasn't just googled "Why do onions make people cry?" the answer is "Some event involving the onion made her cry." Fleshing out the actual event(s) in question can involve either ontological knowledge, as for the onion example, or situational knowledge. For example, if *The raccoon caused the accident*, we know that something that it did caused the accident, which could be anything from running out into traffic to lunging at a dog who then ran into the street. Understanding how a raccoon could cause an accident does not involve knowledge of language, but knowledge of the world. However, before jumping to the conclusion that the need for extensive world knowledge and reasoning represents a dire state of affairs, remember the following: At this point in processing, the LEIA has already detected that an event was elided, and now it is evaluating whether it can easily determine—using ontological knowledge—which one. If it cannot, it will wait until the stage of Situational Reasoning to decide whether it cares. In many cases, it will not. After all, saying *A raccoon caused the accident* might have the discourse function of indicating that it was not the fault of one of the drivers.

5.7 Coreferential Events Expressed by Verbs

Events can be expressed by verbs or by noun phrases. For example, *Someone sneezed loudly* and *There was a loud sneeze* both generate an instance of SNEEZE (LOUDNESS .8) in the TMR. Does the way the event is presented in the text affect how we approach its reference treatment? Yes.

As we have already seen, every noun phrase—whether it instantiates an OBJECT or an EVENT in the TMR—is evaluated with respect to its coreference needs. When *a loud sneeze* is evaluated, the use of the article *a* will block the search for a sponsor. If, by contrast, the input refers to *the loud sneeze*, then the agent will try to identify its sponsor.

Coreference resolution for verbs, however, is approached differently. By default, the agent does not search for a sponsor for each EVENT referred to by a verb. There are two reasons why. First, most verbs introduce new events into the context, which means that the default answer to "Does it have a sponsor?" is "No." Second, when speakers *do* need to establish coreferences among events, they have other ways of doing so apart from using a coreferential verb. We have already seen five of these, repeated below with examples.

(5.110) [A coreferential EVENT can be elided.]
I wanted to kick it but I didn't __. (COCA)

(5.111) [A coreferential EVENT can be expressed by a broad referring expression.]
He didn't even seem sorry and that made me mad. (COCA)

(5.112) [A coreferential event can be expressed using an overt verbal anaphor.]
The focus of debate now is whether Congress will limit the size of the companies or ask the new regulator to do so. (COCA)

(5.113) [A coreferential event can be presented using a full NP. The TMR for this example will contain two instances of RAIN-EVENT with a coreference link.]
Sunday night it rained again, and the rain turned into snow. (COCA)

(5.114) [An event that is a subevent of its sponsor can be presented as a full NP. The ontological description of TRAVEL-EVENT, which is instantiated by *head to*, includes all of the subevents indicated by the underlined NPs.]
And so, decades after Grandma pledged allegiance to the American flag, worn out by my own struggles in the promised land, I heed her ancestral call and head to St. Kitts. The flight is smooth, the landing perfect, the welcome gracious. (COCA)

Of course, one *can* express coreferential events using a sequence of verbs, so the agent has to be on the lookout for such cases. Fortunately, sequences of coreferential verbs often occur in particular types of constructions. Specifically, the clause containing the coreferential verb often occurs immediately after the sponsor clause and includes either an additional modifier (5.115) or the specification of a previously unspecified or underspecified argument (5.116).[47]

(5.115) Mary jogged yesterday, and she jogged so far!

(5.116) Max read all night. He was reading *War and Peace*.

If a clause does not match one of our recorded verbal-repetition constructions, the agent does not seek out a coreferent for its main verb *at this stage*. If there is a coreference relation—either with some event reported in the text or with an event previously remembered by the agent—then this needs to be established during Situational Reasoning, when full reference resolution (i.e., grounding to agent memory) is undertaken.

As a point of comparison, consider how event coreference has been approached in mainstream NLP, where the emphasis has been to support applications involving information extraction, such as template filling, populating databases, and topic tracking. As Lu and Ng explain in their 2018 survey paper, for event mentions to be coreferential they need to be of the same type and have compatible arguments, as in their example:

(5.117) Georges Cipriani {left}$_{ev1}$ a prison in Ensisheim in northern France on parole on Wednesday. He {departed}$_{ev2}$ the prison in a police vehicle bound for an open prison near Strasbourg. (p. 5479)

Lu and Ng explain that event coreferences are more difficult than entity coreferences because (a) establishing event coreferences relies on a larger number of noisy upstream results, including establishing entity coreference relations, and (b) the coreferents can be realized as many types of syntactic categories. All of the works cited by Lu and Ng involve machine learning (ML) since practically all recent work adopts this approach. However, some primarily ML-oriented systems—such as the one reported in Lu and Ng (2016)—do incorporate some handcrafted rules as well.

Lu and Ng (2018) do not sugarcoat the difficulty of the event coreference task or the quite modest capabilities of state-of-the-art systems that they report in a tabular summary of published system evaluations. They explain that many event coreference evaluation scores are artificially high because event mentions are manually identified prior to the evaluation runs—a significant simplification since the identification task (known as *trigger detection*) is itself quite difficult. They summarize, "Both event coreference and trigger detection are far from being solved" (p. 5479). They note that the best results have been achieved for English and that "event coreference models cannot be applied to the vast majority of the world's low-resource languages for which event coreference–annotated data is not readily available" (p. 5484).

Whereas Lu and Ng (2018) focus on system building, Hovy et al. (2013) address a related issue: corpus annotation in service of ML. Their goal is to address some of the more difficult aspects of coreference that have been avoided by past annotation efforts and "to build a corpus containing event coreference links that is annotated with high enough inter-annotator agreement to be useful for machine learning." They establish three levels of event identity: *full identity*; *partial identity*, divided into (a) *membership*, which links multiple instances of the same type of event, and (b) *subevent*, which links events from the same script; and *no identity*. The novel aspect of the work is the second category, which they describe using interesting examples:

> In our work, we formally recognize partial event overlap, calling it *partial event identity*, which permits different degrees and types of event coreference. This approach simplifies the coreference problem and highlights various inter-event relationships that facilitates grouping events into 'families' that support further analysis and combination with other NLP system components. (p. 22)

The latter, as we interpret it, means that the annotations offer a coarse-grained linking of elements of input that require more detailed analysis using methods that are outside the purview of the reported work.

5.8 Further Exploration

1. Read, or at least browse through, "MUC-7 Coreference Task Definition" (Version 3.0, July 1997) by L. Hirschman and N. Chinchor, https://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html. Note how complex the instructions are—how many phenomena have to be explicitly ruled in and ruled out.

2. Explore the state of the art in coreference engines using the Stanford CoreNLP interface at corenlp.run. Be sure to select the coreference annotator by clicking on the Annotations field and selecting "coreference" from the pull-down menu. Experiment with easy inputs and more difficult ones, as described in this chapter.

3. Use the online version of the COCA corpus (https://www.english-corpora.org/coca/) to explore anaphoric event coreference. The most reliable way to identify examples (and avoid false positives) is to search for relevant strings before a "hard" punctuation mark—for example, a period, semicolon, question mark, or exclamation point. Some of the many possible search strings are these (note the spacing conventions required by the interface):

could n't .	would n't .	should n't .	wo 'nt .		
might .	can .	did .	will .	may .	must .
do it .	do that .	did this .	does it .		
has to .	wants to .	tries to .	promised to .		

a. When do these search strings return false positive results (i.e., examples that don't show anaphoric event coreference)? How can they be avoided?

b. What is the sponsor in each example? Can you point to a text string that precisely reflects the sponsor? If not, is the sponsor extralinguistic, is the reference vague, or is something else going on?

c. If there is a textual sponsor, is there type or instance coreference of the verbs? Type or instance coreference of the internal arguments? Does the sponsor clause include modal or other scoping verbs? If so, are they included in or excluded from the resolution?

d. Is each example easy or difficult for a system to automatically resolve? Why?

4. Looking just at the table of contents at the beginning of the book, try to reconstruct what was discussed in each section of chapter 5 and recall or invent examples of each phenomenon.