Linguistics for the Age of AI

OA Funding Provided By:

The open access edition of this book was made possible by generous funding from Arcadia—a charitable fund of Lisbet Rausing and Peter Baldwin.

The title-level DOI for this work is:

doi:10.7551/mitpress/13618.001.0001

Notes

Chapter 1

1. For historical overviews of machine translation, see, e.g., Hutchins (1986) and Nirenburg et al. (2003). Portions of this discussion were originally published as Nirenburg & McShane (2016a).

2. A similar approach is a cornerstone of the current corpus-based paradigm in NLP, which usually involves the analysis not of text meaning in toto but either the treatment of the meaning of selected textual strings or no treatment of meaning at all—instead orienting exclusively around the syntactic and morphological properties of words.

3. This survey was published in Bar Hillel (1970).

4. For overviews of head-driven phrase structure grammar and lexical functional grammar—along with generative grammar and construction grammar—see Pustejovsky & Batiukova (2019, Chapter 3).

5. The automatic learning of knowledge by the agent has also started to be addressed in cognitively inspired approaches to multifunctional agent modeling. See, e.g., Forbus et al. (2007), Navigli et al. (2011), Nirenburg et al. (2007), and Wong et al. (2012). See the deep dive in section 1.6.2 for additional details about learning by LEIAs.

6. The approach to recording knowledge and computing meaning that we will describe was originally developed for natural language processing outside of a full agent architecture (Nirenburg & Raskin, 2004) but has easily accommodated agent-oriented extensions.

7. There have been ongoing debates between eliminativists and functionalists on the status of unobservables but, for our needs, a human-level explanation of behavior cannot be formulated in terms of neuronal activity. We prefer possibly "naive" explanations that may ultimately not be true in scientific terms but are expected to be accepted as explanations by regular people in regular circumstances, outside a philosopher's study or a psychologist's lab.

8. Although nonlinguistic channels of perception are largely outside the scope of this book, we have been working on them in earnest and include select mentions. For example, chapter 8 describes interoception by virtual patients (section 8.1.3.2) and vision by physical robots (sections 7.7.2 and 8.3), both modeled within the OntoAgent architecture.

9. For an analysis of how annotated corpora can support the development of linguistic theories, see de Marneffe & Potts (2017).

10. Readers of our draft manuscript offered interesting suggestions for additions to this literature review. However, we kept this review, like others throughout the book, highly selective in order not to stray from the main narrative.

11. FrameNet, described later in the chapter, is a computationally oriented resource deriving from early work in construction grammar led by Charles Fillmore.

12. This theory was actually used by Purver et al. (2011) as the theoretical substrate for an incremental parser. However, although that parser included semantics, it appears to be a version of upper-case semantics—i.e., the intended interpretations were provided manually.

13. See Chambers et al. (2002) for experiments exploring how the affordances of objects in a workspace affect subjects' interpretations of language inputs. For reviews of the literature, see Kruijff et al. (2007) and Chambers et al. (2004).

14. For a discussion of various approaches to semantic analysis in cognitive systems, see McShane (2017a).

15. Thesauri are quite diverse. To take two extremes, Apresjan's (2004) *New Explanatory Dictionary of Russian Synonyms* was manually created, is extremely rich in detail, and was intended for use by people. By contrast, Inkpen & Hirst's (2006) lexical knowledge base of near-synonym differences was compiled automatically, using a multistage process of machine learning, and was intended for use by systems.

16. Manning (2006) promotes the idea of bridging work between the NLP and Knowledge Representation and Reasoning (KR&R) communities, writing: "NLP people can do robust language processing to get things into a form that KR&R people can use, while KR&R people can show the value of using knowledge bases and reasoning that go beyond the shallow bottom-up semantics of most current NLP systems."

17. A good example is Leafgren's (2002) description of how different kinds of referring expressions are used in Bulgarian.

18. A recent contribution on the resource front is the Stanford Natural Language Inference corpus, which includes 570,152 labeled pairs of captions (Bowman et al., 2015). Users of the Amazon Mechanical Turk environment were presented with image captions—without the images themselves—and asked to write three alternative captions: one that was definitely true (an entailment), one that may or may not be true (a neutral statement), and one that was definitely false (a contradiction). The task instructions included explanations of the quartet:

> Two dogs are running through a field. (prompt)
> There are animals outdoors. (entailment)
> Some puppies are running to catch a stick. (neutral)
> The pets are sitting on a couch. (contradiction)

This annotated corpus has been used for experiments in machine learning related to natural language inference. We do not think that this corpus will facilitate knowledge acquisition for agent systems due to the unconstrained selection of *any* entailment, *any* neutral statement, and *any* contradiction.

19. Past theoretical work linking cognitive linguistics and computation includes, e.g., Feldman & Narayanan (2004) and Feldman (2006).

20. This short survey did not even touch on the field of neurolinguistics because agent development does not attempt the biological replication of a human brain, and it remains to be seen whether and how the results of neurolinguistics will ever inform computational cognitive modeling.

21. We do not include here the resources that will be described in upcoming chapters.

22. For a discussion of commonsense reasoning and knowledge, see Davis & Marcus (2015).

23. Note that research and development in the area of component integration has become increasingly prominent (e.g., Bontcheva et al., 2004; Shi et al., 2014). An additional benefit of the integrative approach is the possibility of viewing components of application systems as black boxes communicating with other components only at the input-output level, thus allowing integration of components built using potentially very different approaches.

24. Much thought has been given to understanding and repairing this state of affairs; see, e.g., Clegg & Shepherd (2007) for an analysis within the biomedical domain.

25. Paraphrase detection has become a well-investigated topic in itself; see Magnolini (2014) for a survey.

26. See S. Clark (2015) for an excellent overview, including historical references.

27. Work is underway to extend distributional semantics to exploit compositionality (Goyal et al., 2013).

28. A detailed discussion of phenomenology and its influence on our view of modeling agents is outside the scope of this book. See Zlatev (2010), Löwe & Müller (2011), Carruthers (2009), and Andler (2006) for relevant discussions.

29. This learning capability is realistic only when the agent is already endowed with a critical mass of knowledge of the world and language. In our work, we do not address early development stages. We concentrate on modeling the behavior of adult humans.

30. Comprehensive LEIA-based applications address other types of learning as well, such as the learning of ontological scripts (Nirenburg et al., 2018).

31. These ideas were first published in McShane (2017b).

32. "Detroit," Wikipedia, accessed December 10, 2016, https://en.wikipedia.org/wiki/Detroit

33. Some annotations have been removed for concise presentation.

34. To be done properly, such comparisons require (a) a near-developer's understanding of each environment, which is hardly ever achievable using published materials because the environments are constantly in flux; (b) a thoughtful, selective process of analysis, which could consume a level of effort equivalent to several doctoral theses; and (c) a presentation strategy that provides readers with all the necessary background about each environment referred to, which would constitute a separate book.

35. Further reading includes Stich & Nichols (2003) for folk psychology; Malle (2010) for attribution theories; Bello (2011) for mindreading; and Carruthers (2009) for metacognition.

36. An extended version of this work is reported in DeVault et al. (2011).

37. Throughout the book, examples drawn from the COCA corpus (Davies, 2008–) will be indicated by the subscript (COCA) following the example.

38. For a list of coreference annotation schemes for NPs, see Hasler et al. (2006).

39. Here we will focus on English, but the same principles apply to other language/culture pairs.

40. Dialog acts have also been called speech acts; locutionary, illocutionary, and perlocutionary acts; communicative acts; conversation acts; conversational moves; and dialog moves. See Traum (1999b, 2000) for references to the literature.

41. Another sample classification is found in Traum (1994, p. 57), which identifies four conversation act types that are relevant for different-sized chunks of conversation. Traum presents each with a sample of the associated dialog acts, as shown below (we normalize capitalization and spell out abbreviations):

turn-taking:	Take-turn, Keep-turn, Release-turn, Assign-turn
grounding:	Initiate, Continue, Ack[nowledgment], Repair, Req[uest]Repair, Req[uest] Ack[nowledgment], Cancel
core speech acts:	Inform, YNQ [i.e., ask question], Check, Eval[uate], Suggest, Request, Accept, Reject
argumentation:	Elaborate, Summarize, Clarify, Q&A, Convince, Find-Plan

42. The formatting of dialog act names (small caps, capitalization conventions, and singular vs. plural) is directly from Stolcke et al.'s (2000) table 2 (p. 341), which also includes corpus examples of each dialog act and its corpus-attested frequency.

43. For related literature see, e.g., the special issues of *Robotics and Autonomous Systems* (Coradeschi & Saffiotti, 2003) and *Artificial Intelligence* (Roy & Reiter, 2005), as well as Roy (2005), Scheutz et al. (2004), Gorniak & Roy (2005), and Steels (2008).

44. However, in field-wide competitions that pit knowledge-lean systems head-to-head, annotated corpora are often provided not only for the training portion but also for the evaluation portion. This means that such systems are actually not operating exclusively over observable data.

45. The Message Understanding Conferences (MUCs; Grishman & Sundheim, 1996) and Text Retrieval Conferences (TRECs; http://trec.nist.gov/) are noteworthy examples.

46. For surveys of the literature, see Lapata (2002), Girju et al. (2005), Lieber & Štekauer (2009), and Tratz & Hovy (2010).

47. Similarly, Lapata (2002) developed a probabilistic model covering only those two-noun compounds in which N1 is the underlying subject or direct object of the event represented by N2: e.g., *car lover*.

48. Accessed June 6, 2020, https://ufal.mff.cuni.cz/pcedt2.0/publications/t-man-en.pdf.

Chapter 2

1. Simplifications include removing most of the metadata that TMRs typically carry. We will not comment further on the specific simplifications of the different knowledge structures to be presented throughout.

2. Upper-case semantics refers to the practice, undertaken by some researchers in formal semantics and reasoning, of avoiding natural language challenges like ambiguity and semantic non-compositionality by

asserting that strings written using a particular typeface (often, uppercase) have a particular meaning: e.g., TABLE might be said to refer to a piece of furniture rather than a chart.

3. The TIME slot includes a call to a procedural semantic routine, *find-anchor-time*, that can attempt to concretize the time of speech if the agent considers that necessary. In many applications, it is not necessary: the general indication of past time—i.e., "before the time of speech"—is sufficient.

4. At the time of writing we are using the Stanford CoreNLP toolset (Manning et al., 2014). This toolset was updated several times during the writing of this book. The most recent version (version 4.0.0) was released late in the book's production process. We made an effort to update examples and associated discussions accordingly. We present screen shots of CoreNLP output for some examples in the online Appendix at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html.

5. A more complete lexicon would include many more constructions, such as *eat one's hat*, *eat one's heart out*, *eat someone alive*, and so on.

6. One might ask, Why is there no concept for EAT in the ontology? This decision reflected the priorities and goals of knowledge acquisition at the time this corner of the ontology/lexicon pair was being acquired.

7. LEIAs also use several auxiliary knowledge bases, rule sets, and algorithms, which will be described in conjunction with their associated microtheories.

8. There has been much debate about the optimal inventory of case roles for NLP systems. Some resources use a very large inventory of case roles. For example, O'Hara & Wiebe (2009) report that FrameNet (Fillmore & Baker, 2009) uses over 780 case roles and provides a list of the most commonly used 25. Other resources underspecify the semantics of case roles. For example, PropBank (Palmer et al., 2005) uses numbers to label the case roles of a verb: Arg0 and Arg1 are generally understood to be the agent and theme, but the rest of the numbered arguments are not semantically specified. This approach facilitates the relatively fast annotation of large corpora, and the resulting annotations support investigation into the nature and frequency of syntactic variations of the realization of a predicate; however, it does not permit automatic reasoning about meaning to the degree that an explicit case role system does.

9. For early work on scripts, see Minsky (1975), Schank & Abelson (1977), Charniak (1972), and Fillmore (1985).

10. The scientific question related to this engineering solution for writing scripts is, Can we write all of this knowledge using a nonprogramming-oriented, static formalism (something that looks more typically ontological) and then write a program that automatically generates code to drive the actions of the agent? Phrased differently, can we write programs to write programs? We will leave this problem to the field of automatic programming or source-code generation. It is a fascinating problem whose full exploration would take our core research program too far afield. Instead of pursuing this, our knowledge engineers write scripts using semiformal knowledge representation strategies that include tables, slot-filler structures, and even diagrams (see chapter 8 for examples), and then they directly collaborate with the programmers who engineer all the required system behavior.

11. McShane et al. (2016) detail that NLU system.

12. For approaches to incremental syntactic parsing see, e.g., Ball (2011) and Demberg et al. (2013).

13. The Pens are the Pittsburgh Penguins, a professional ice hockey team.

14. This dichotomy might be considered mutatis mutandis, parallel to Newell's (1982) distinction between the knowledge level and the symbol level(s) in AI systems.

15. Discussing this very complex issue in any detail is beyond the scope of this book. For an introduction to the relevant topics, see Frigg & Hartman (2020) and Winther (2016).

16. Our conception of a model is strategically congruent with the views of Forbus (2018, Chapter 11), though we concentrate on modeling less observable phenomena than those in the focus of Forbus's presentation. Our views on the nature of theories, models, and systems have been strongly influenced by Bailer-Jones (2009), but for the purposes of our enterprise we do not see a need to retain the same fine grain of analysis of these concepts and their interrelationship as the predominantly philosophy-of-science angle of Bailer-Jones.

17. The psycholinguistic evidence that Wilks and Ide cite to support this position is irrelevant because systems do not operate the way people do. In fact, Wilks's own famous "theorem" to the effect that there is no linguistic theory, however bizarre, that cannot be made the basis of a successful NLP system (Wilks actually said "MT system") seems to argue for discounting psycholinguistic evidence for NLP.

18. Border collies are the real geniuses, with a default INTELLIGENCE value of 1.

19. For a comparison of upper ontologies see Mascardi et al. (2007) and references therein.

20. In reviewing the utility of a number of large medical ontologies for NLP, Hahn et al. (1999) report that MeSH shows "semantic opacity of relations and concepts … [and a] lack of formal concept definitions"; in SNOMED, "only a small part of all possible [automatically generated] combinations of axes correspond to consistent and reasonable medical concepts"; and in UMLS, "by merging concepts collected from different sources, a problematic mixture of the semantics of the original terms and concepts is enforced." It should be mentioned that these resources, like the very widely used WordNet, were actually developed for people, not NLP.

21. We believe that history will look back on this period of building wordnets with similar disillusionment, at least with respect to sophisticated NLP applications, for which disambiguation is crucial.

22. As Wilks et al. (1996, pp. 121–122) describe the situation, many practitioners of NLP consider lexicons "primarily subsidiary" resources expected to "fall out of" their theory of choice.

23. The languages are Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. The work continues the earlier PAROLE project, which developed 20,000-sense morphological and syntactic lexicons for these languages.

24. The notion of a largely language independent lexicon was introduced in McShane et al. (2005a).

25. If, for example, the first LEIA lexicon was created for German, it is possible that the acquirer would create a special concept for WHITE-HORSE, since German has a word for that, *Schimmel*. If an English lexicon were bootstrapped from the German one, the acquirer would either ignore that lexical sense and concept or encode the multiword expression *white horse* to map to it.

26. We could, alternatively, have created a concept WEAPON-OF-MASS-DESTRUCTION, whose parent would be WEAPON and whose children would be CHEMICAL-WEAPON and BIOLOGICAL-WEAPON. We tend not to create special concepts for groups of things, however, since there are expressive means to group things, as needed, in the lexicon.

27. There has been extensive work on linguistic paraphrase in the knowledge-lean paradigm which involves, e.g., determining the distributional clustering of similar words in corpora (e.g., Pereira et al., 1993; Lin, 1998), using paraphrases for query expansion in question-answering applications (e.g., Ibrahim et al., 2003), and automatically extracting paraphrases from multiple translations of the same source text (Barzilay & McKeown, 2001).

28. One could, of course, create a multiword lexical sense for *go (somewhere) by plane*, which would map to AERIAL-MOTION-EVENT and avoid this particular case of ontological paraphrase. However, the point of this example is to show how ontological paraphrase can be reasoned about when it does occur.

Chapter 3

1. See de Marneffe et al. (2006) for a description of generating dependency parses from phrase structure parses.

2. Captured November 8, 2019. Figures 3.1 and 3.2 originally appeared in color.

3. Recall that ^, used in the sem-struc zones of lexicon entries, indicates *the meaning of.*

4. For early work on noncanonical input, see Carbonell & Hayes (1983).

5. Some annotations have been removed to keep the presentation concise.

6. You might wonder if we are making the task artificially more difficult by providing LEIAs with only written transcripts instead of the speech stream itself—after all, speech includes prosodic features that assist people in extracting meaning. No doubt, prosodic features could be very useful to LEIAs; however, the precondition for using them remains outstanding. Specifically, methods must be developed to automatically extract and interpret such features within the agent's ontological model.

7. There are formalism-related reasons why we cannot have a single template that would encompass all three.

8. For further description of this approach, albeit within a different implementation of the NLU system, see McShane et al. (2016).

9. For purposes of clarity, we focus on verbs as argument-taking heads, though other parts of speech can take arguments as well.

10. The LEIA *can* detect certain kinds of plays on idioms, but not at this stage; that occurs later on, during Extended Semantic Analysis.

11. This output was confirmed on January 25, 2020, using the online interface at the website corenlp.run.

12. Note that the example intentionally includes no coreferential expressions (e.g., no definite articles), so we cannot assume that the preceding linguistic context will aid in disambiguation.

13. For past work on semantic sense bunching and underspecification, see, e.g., Buitelaar (2000) and Palmer et al. (2004).

14. We use various sense-numbering conventions in the lexicon for internal bookkeeping.

Chapter 4

1. At the time of writing, we are working on a project whose objective is to record the algorithms and code base of not only our NLU system but also the OntoAgent cognitive architecture overall, using the graphic representations of the Unified Modeling Language™ (UML). Future publications reporting the results of that work will provide interested developers with the algorithms actually implemented in our current NLU engine.

2. Consider the following distinction, which might be missed by all but the most informed foodies: "Note the difference between smoked Scottish salmon and Scottish smoked salmon. It's possible that the wording of the latter has been deliberately used to cover the fact that the salmon has been smoked in Scotland, but not necessarily sourced from Scotland." From "Scottish or Scotch? A Guide to Interpreting Food Labels," The Larder, *The List*, May 1, 2009, https://food.list.co.uk/article/17265-scottish-or-scotch-a-guide-to-interpreting-food-labels/.

3. Garden-path sentences are grammatical sentences for which the hearer's initial interpretation of the first part ends up being incorrect when the sentence is completed. A classic example is *The horse raced past the barn fell*.

4. For earlier work on adjectives in this paradigm, see Raskin & Nirenburg (1998).

5. The "sad" meaning of *blue house* and the "colored blue" meaning of *blue person* would not be generated because they are instances of nonliteral language. Nonliteral interpretations are generated only if there is a trigger to do so, such as an incongruity during processing (see section 6.2), which is not the case here.

6. We do not adopt procedural semantics as our overarching lexical theory. That is, our approach would not associate common objects like *chair* with procedural routines to seek out their extensions (see Wilks et al., 1996, pp. 20–22 for a discussion of the various interpretations of "procedural semantics"). Instead, we apply the term to only that subclass of lexical phenomena that predictably requires a context-bound specification of their meaning, for which the lexicon explicitly points to the necessary routine.

7. *Evaluative* is actually a type of modality. The treatment of modality and aspect is described in sections 4.2.1 and 4.2.2. For clarity of presentation in this set of examples, we use a shorthand representation for the full modality and aspect frames.

8. If we were to use proper names instead, their TMR frames would be instances of HUMAN described using the feature HAS-PERSONAL-NAME. The TMRs omit indications of time.

9. SET-1 and SET-2 are generated by a lexical sense for the conjunction *and* in which *and* requires two noun phrases as its arguments.

10. We chose an example that describes the set as participating in an event because this most clearly illustrates why set expansion is needed: the agent must understand that each of the set elements is engaged in its own instance of the event. However, even if no event is mentioned, this type of set expansion best serves downstream reasoning. *Two gray wolves* generates two different instances of WOLF, each of which is described by "COLOR gray."

11. We do not present the formal meaning representations, since they are representationally complex and are not needed to convey our main point, which is conceptual.

12. It is stripped of metadata and the call to resolve the reference of *the*.

13. See Nirenburg & Raskin (2004) for motivation and a more formal description of this inventory.

14. One can also demand an action that is not agentive, such as when the interlocutor will be the experiencer: *Get better soon!* or *Catch the flu so you can get out of your final!*

15. According to Monti et al. (2018, p. 3), "Biber et al. (1999) argue that they [multiword expressions] constitute up to 45% of spoken English and up to 21% of academic prose in English. Sag et al. (2002) note that they are overwhelmingly present in terminology and 41% of the entries in WordNet 1.7 are reported to be multiword units."

16. We also do not adopt most of the theoretical principles they advocate, such as the lack of empty categories, the lack of transformations, and an inheritance network of constructions. But those aspects are not directly related to the definition of the word *construction*, which is our interest here.

17. Our approach to integrating all kinds of constructions into the lexicon resonates with Stock et al.'s (1993, p. 238) opinion that idioms (a subtype of constructions) should be integrated into the lexicon as "more information about particular words" rather than treated using special lists and idiosyncratic procedures.

18. Other names for multiword expressions, and subtypes of them, are *multiword units*, *fixed expressions*, *set expressions*, *phraseological units*, *formulaic language*, *phrasemes*, *idiomatic expressions*, *idioms*, *collocations*, and *polylexical expressions* (Monti et al., 2018, p. 3). Monti et al. provide a reference-rich overview of work on multiword expressions in language processing technologies, with an emphasis on machine translation.

19. For more extensive literature reviews, see Cacciari & Tabossi (1993) or McShane, Nirenburg, & Beale (2015). The analysis presented here draws from the latter.

20. This section draws from McShane et al. (2014).

21. Tratz & Hovy (2010) prioritized achieving interannotator agreement in a Mechanical Turk experiment, and this methodology influenced their final inventory of relations. Thanks to them for sharing their data.

22. This is just one of many possible control strategies, another being to launch all analysis functions, score them, and select the one with the highest score.

23. This section draws from Nirenburg & McShane (2016b).

24. The mid-2010s gave rise to a new wave of research on metaphor in the NLP community that follows the accepted knowledge-lean paradigm. But, as Shutova (2015) writes, "So far, the lack of a common task definition and a shared data set have hampered our progress as a community in this area. This calls for a unification of the task definition and a large-scale annotation effort that would provide a data set for metaphor system evaluation …" (p. 617). Whether that will occur is one question, and whether it will actually address the full complexity of metaphors is another.

25. Quoted from "The D Team" by Lee Edwards, *American Spectator*, April 30, 2013, https://spectator.org/33713_d-team/

26. A more comprehensive, but also more complex, approach would be to posit a disjunctive set allowing for either an event or a state.

Chapter 5

1. This process is called *grounding* in the robotics community. *Grounding* has a different meaning in the area of discourse and dialog processing: it refers to indicating—using utterances (e.g., *uh-huh*) or body language (e.g., nodding)—that one has understood what the other person said and meant.

2. There is no bucket involved in the idiomatic meaning of *kick the bucket*. There is also no kicking in the usual sense; however, when used in this idiom, *kick* means DIE, and DIE is referential.

3. This section draws from McShane (2009).

4. Work on associated topics has been carried out in the knowledge-lean paradigm. Boyd et al. (2005), Denber (1998), Evans (2001), and Li et al. (2009) describe methods for automatically detecting pleonastic *it*; Vieira & Poesio (2000) present a system for automatically processing definite descriptions; Bean & Riloff (1999) describe a method for identifying nonanaphoric noun phrases.

5. In the TMR, this two-part idiom will instantiate the property CONTRAST, whose domain is the meaning of the first proposition and whose range is the meaning of the second.

6. In the TMR, PLUMBER will fill the property HAS-SOCIAL ROLE, as follows: HUMAN-1 (HAS-PERSONAL-NAME 'Danny') (HAS-SOCIAL-ROLE PLUMBER).

7. *Has* is used here as an auxiliary verb, not a main verb.

8. The study reported in Elsner & Charniak (2010) shows that, given a coreference window of ten sentences using the MUC-6 coreference dataset, only about half of same-headed NPs were coreferential. However, this is a very large window of coreference, making this result not entirely surprising.

9. See McShane (2005, Chapter 7) for a discussion of such phenomena crosslinguistically.

10. For discussions of bridging see, e.g., Asher & Lascarides (1996) and Poesio, Mehta, et al. (2004).

11. Recasens et al. (2012) refer to some of these as near-identity relations and discuss their treatment in corpus annotation.

12. For LEIAs, the ambiguity of universally known definite descriptions is accounted for by multiple lexical senses. For example, one sense of *sun* requires *the* and means 'the Earth's sun,' and another sense does not require *the* and can indicate the sun of any planet. This is just another matter of lexical disambiguation.

13. See, e.g., Marcu (2000).

14. Some approaches orient around utterances rather than sentences.

15. For a discussion of treating ambiguity in corpus annotation, see Poesio & Artstein (2005).

16. For example, intransitive verbs like *sleep* require only a subject; transitive verbs like *refinish* require a subject and a direct object; and ditransitive verbs like *give* require a subject, a direct object, and an indirect object.

17. Bos & Spenader (2011) summarize NLP's avoidance of ellipsis as follows: "First, from a purely practical perspective, automatically locating ellipsis and their antecedents is a hard task, not subsumed by ordinary natural language processing components. Recent empirical work (Hardt, 1997; Nielsen, 2005) indeed confirms that VPE [verb phrase ellipsis] identification is difficult. Second, most theoretical work begins at the point at which the ellipsis example and the rough location of its antecedent are already identified, focussing on the resolution task" (pp. 464–465). The field of generative syntax, for its part, has studied ellipsis in earnest for decades but, in accordance with its purview and goals, it focuses exclusively on syntax and licensing conditions, not on semantics or resolution.

18. The CoreNLP developers call this tool dcoref, for deterministic coreference annotator. Other external coreference resolution systems could be used if they were to offer better quality of results or coverage of phenomena. In general, we welcome imported solutions as long as the overhead of incorporating them is not prohibitive.

19. Poesio et al. (2016, pp. 88–90) also provide a nice overview, within a broader description of the field. For an example of a contribution that integrates more knowledge into a still primarily machine-learning approach, see Ratinov & Roth (2012).

20. The CoreNLP output for several examples in this chapter, including (5.35) and (5.36), is presented in the online appendix at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html.

21. The fact that we would more naturally verbalize the sponsor of *it* as "her being happy" or "the fact that she is happy" is inconsequential. Coreferences are actually established at the level of semantic interpretations (TMRs), not text strings.

22. The constructions described here were actually inspired by McShane's work on argument ellipsis in Russian and Polish (McShane 2000, 2005). In these languages, the ellipsis of referential direct objects is permitted only with substantial linguistic or contextual support, and that often involves parallelism. Not surprisingly, some of the same lexico-syntactic constructions that permit argument ellipsis in Russian and Polish quite confidently predict the coreference of overt arguments in English. This overlap is not only of theoretical interest; it also suggests that, viewed crosslinguistically, a knowledge-based approach to reference resolution is both feasible and economical in terms of the descriptive work needed.

23. This evaluation excluded *it* because we did not leverage semantic analysis for the evaluation and, therefore, the system could not detect pleonastic and idiomatic usages.

24. We tasked the system only with identifying the nominal head of the antecedent, not the entire noun phrase, which might include a determiner, adjectives, or relative clauses. The full NPs are indicated in the examples for clarity's sake.

25. Developing a comprehensive analysis system means that not all components are necessarily ready to go at a given time. Some of our evaluations have involved only a subset of overall system capabilities, as discussed in chapter 9.

26. *Span of text* is a syntactic description. Semantically, we are talking about multiple propositions.

27. For example, *Horatio proposed that everyone in the office should go on a group jog every morning, but his suggestion was met with collective horror.*

28. Byron (2004) is an exception but has narrow domain coverage. (It provides a nice review of the linguistic literature on broad referring expressions.) Machine translation systems must treat broad RefExes, but they have the advantage of not having to actually resolve them: they can replace a vague expression in source language by an equally vague one in the target language.

29. Parentheses indicate optionality, a forward slash indicates a choice, caps indicate category types, and underlining indicates coreferential categories.

30. The evaluation (reported in McShane & Babkin, 2016a) included twenty-seven contexts, for which twenty-five answers were correct, one was incorrect, and one was partially correct.

31. Past research has shown that discourse segments that serve as sponsors for broad RefExes are almost always contiguous with the broad RefEx's clause—i.e., discourse segments are not skipped over. See Byron (2004) for a review of that literature.

32. More formally, simple syntactic structures have none of the following dependencies in the CoreNLP dependency parse: *advcl*, *parataxis*, *ccomp*, *rcmod*, *complm*, *dep*, *conj* (with verbal arguments), *xcomp* (with a lexically recorded matrix verb as the governor), or *aux* (not involving a tense marker). We do not assume that all CoreNLP dependency parses will be error-free; however, its accuracy for detecting syntactically simple clauses using this algorithm is quite good.

33. For more on sentence trimming in the literature, see the references cited in McShane, Nirenburg, & Babkin (2015).

34. The full list is available at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html.

35. For work on near-synonyms, see DiMarco et al. (1993) and Inkpen & Hirst (2006).

36. The first stage of work involved compiling a test list of verbs for which either the subject or the object was narrowly constrained, then compiling a list of typical fillers for that role. We used a total of 202 verbs with an average of nearly 60 keywords each. However, the average was pulled up by verbs like *eat/cook* and *die*, for which hundreds of food items and animals, respectively, were listed as keywords. For details, see https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html. We used the Gigaword corpus (Graff & Cieri, 2003) for the experiment.

37. Formally speaking, these descriptions unify.

38. We know of no definitive, comprehensive list of such entities. Automatically generated lists of this kind include numerous false positives.

39. Of course, metaphorical and other senses can be recorded as well.

40. By contrast, nonrestrictive modifiers—which are usually set off by commas—provide additional information but are not as essential to the meaning of the sentence: e.g., *Our neighbors' dog, Captain, comes over often*.

41. Accessed January 1, 2019, https://en.wikipedia.org/wiki/Lion.

42. HAS-OBJECT-AS-PART and PART-OF-OBJECT are inverse relations in the ontology.

43. The naming conventions for concepts are irrelevant; they are for human orientation only. The agent understands their meaning as the set of property fillers defined in the ontology.

44. As regards past work on VP ellipsis outside our group, Johnson (2001) offers a descriptive account aptly entitled "What VP ellipsis can do, what it can't, but not why." In the computational realm, if knowledge-lean systems treat VP ellipsis at all, they do not address the semantic issues and they do not offer confidence estimates for resolutions. Hardt (1997) reports a system for resolving VP ellipsis that required a manually corrected parse and did not pursue semantics. Most work on instance versus type coreference of internal arguments has been carried out in the paradigm of theoretical linguistics, which does not offer heuristics that can guide system building.

45. AECs can also be expressed using pronouns and other descriptions, but those cases are discussed in their respective sections.

46. The terms *strict* and *sloppy* coreference are used in the generative syntax literature (see, e.g., Fiengo & May, 1994) to refer to objects showing what we call *instance* and *type* coreference.

47. See McShane (2005) for further discussion of the reference-oriented effects of repetition structures.

Chapter 6

1. Recall that many procedural semantic routines involve coreference and were resolved during Basic Coreference Resolution.

2. This dovetails with the views of Lepore & Stone (2010) about metaphor—that is, metaphorical meanings do not need to be fully semantically interpreted or recorded.

3. This description may bring to mind the statistical approach called distributional semantics. Distributional semantics operates over uninterpreted text strings, not meanings, and therefore (a) it is very noisy due to lexical ambiguity, and (b) it does not yield formal representations to support the agent's downstream reasoning.

4. COLOR is a literal attribute, which means that its values are not ontological concepts. That is why they are not in small caps.

5. Recall that ontological scripts use a wide variety of knowledge representation methods, far beyond the simple slot-facet-filler formalism used for the nonscript portion of the ontology.

6. Section 7.9 describes automatic learning by reading, which can be useful for this kind of knowledge acquisition.

7. This is different from the property RELATION, which is the head of a large subtree of the ontology.

8. See Onyshkevych (1997) for a discussion of shortest-path algorithms.

9. For reasoning by analogy see, e.g., Forbus (2018), Gentner & Smith (2013), and Gentner & Maravilla (2018).

10. The idea of a metonym repository ascends at least to Fass (1997).

11. The linguistic literature has identified several other types of performance errors as well, such as spoonerisms (saying *The Lord is a shoving leopard* instead of *The Lord is a loving shepherd*) and cases of anticipation (saying *bed and butter* instead of *bread and butter*). These are not particularly common, so their treatment is not a priority.

12. For a book-length cognitive model of idiomatic creativity, along with extensive manual analyses of data, see Langlotz (2006). This model does not directly inform our work because it is not a *computational* cognitive model—it lacks heuristics that would make the human-oriented observations automatable. For an entertaining take on linguistic creativity overall, see Veale (2012).

13. In the lexicon, these will be recorded using the more involved formalism of that knowledge base. We use shorthand here for readability's sake.

14. Although theoretically oriented accounts have classified sources of idiomatic creativity—e.g., Langlotz's (2006, pp. 176–177) model distinguishes *institutionalized variants*, *occasional variants*, *pun variants*, and *erroneous variants*, among others—that grain size of analysis is unrealistic for computational models within the current state of the art.

15. This example uses generic *you*, which we refer to here, for simplicity's sake, as an instance of HUMAN.

16. Recall that many procedural semantic routines were already run by this time, during both Basic Semantic Analysis and Basic Coreference Resolution.

17. K. B. Cohen et al. (2008) present a rigorous linguistic analysis in a related vein. They studied the alternations in the argument structure of verbs commonly used in the biomedical domain, as well as their associated nominalizations. The work was aimed at improving information extraction systems.

18. Bakhshandeh et al. (2016) divide up linguistic phenomena related to comparisons and ellipsis differently than we do, and they treat them using supervised machine learning. Although their target knowledge structures are deeper than those used by most machine learning approaches, their direct reliance on annotated corpora, unmediated by a descriptive microtheory, makes their results not directly applicable to our work.

19. A simple example of leaving a procedural semantic routine unresolved involves the representation of the past tense. Every past-tense verb spawns a TMR whose TIME slot is filled with the call to the procedural semantic routine '<*find-anchor-time*', meaning 'before the time of speech'. If this procedure is run, the agent attempts to determine the time of speech/writing, which may or may not be available and may or may not matter.

20. Some points of comparison with the literature are as follows. Fernández et al. (2007) identify fifteen classes of what they call "non-sentential utterances" (NSUs), which they use in their work on automatically classifying NSUs using machine learning. Cinková (2009) describes an annotation scheme for detecting and reconstructing NSUs. And Schlangen & Lascarides (2003) identify twenty-four speech act types that can be realized with NSUs, grounding their taxonomy in the rhetorical relations defined by a theory of discourse structure called SDRT (Asher, 1993; Asher & Lascarides, 2003). Although some of the descriptive work and examples from these contributions can inform the development of our microtheory of NSUs, the goals pursued are so different that comparisons are quite distant. Cinková's corpus annotation work is targeted at developers pursuing supervised machine learning. Fernández et al.'s approach assumes a downstream consumer as well: "Our experiments show that, for the taxonomy adopted, the task of identifying the right NSU class can be successfully

learned, and hence provide a very encouraging basis for the more general enterprise of fully processing NSUs." As for Schlangen & Lascarides's approach, it relies on Minimal Recursion Semantics (Copestake et al., 2005), a minimalistic approach to lexical and compositional semantic analysis, which they describe as "a language in which partial descriptions of formulae of a logical language (the *base language*) can be expressed. This allows one to leave certain semantic distinctions unresolved, reflecting the idea that syntax supplies only partial information about meaning. Technically this is achieved via a strategy that has become standard in computational semantics (e.g., (Reyle, 1993)): one assigns labels to bits of base language formulae so that statements about their combination can remain 'underspecified' " (p. 65).

21. This work was originally described in McShane et al. (2005b).

22. See Ginzburg & Sag (2001) for more on interrogatives.

Chapter 7

1. Bello & Guarini (2010) discuss mindreading as a type of mental simulation.

2. Sample problems are available at https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.xml

3. Some others require textual coreferences that the agent has not yet been able to identify. This can happen, e.g., for difficult uses of demonstrative pronouns.

Chapter 8

1. In the realm of medical pedagogy, *virtual patients* have also been defined as physical manikins, as live actors who role-play with trainees, and as computer programs that allow trainees to work through static decision trees.

2. See Bailer-Jones (2009) for a discussion of modeling within the philosophy of science.

3. The abstract features used for cognitive modeling are similar to the intermediate categories in ontologies. Although regular people might not think of WHEELED-AIR-VEHICLES as a category, this can still be an appropriate node in an ontology.

4. See section 2.3.1 for an overview of how scripts are represented in the LEIA's ontology.

5. This description draws from McShane, Nirenburg, et al. (2007).

6. See section 4.1.1 for a discussion of property classes and their allowable values.

7. For a description of emotional effects on GERD, see Mizyed et al. (2009). For our incorporation of these factors into the clinical model, see McShane, Nirenburg, Beale, et al. (2013).

8. See section 2.8.2 for a discussion of why a direct mapping between a word and an ontological concept does not constitute upper-case semantics.

9. See section 2.3.1 for a discussion of inheritance in the ontology.

10. See section 4.1.1 for examples of scalar attributes.

11. Remember, diseases can also be dynamically generated if their preconditions are met.

12. This is an arbitrary large number that signifies "never."

13. For a review of text meaning representations, see the warm-up example in section 2.2.

14. A more detailed version of this analysis was published as McShane, Nirenburg, Jarrell, & Fantry (2015).

15. Analogous difficulties are well-known in the domain of ontology merging.

16. See section 2.8.4 for a discussion of paraphrase.

17. As Cooke (n.d.) reports, reviews and categorization schemes for knowledge elicitation and modeling "abound." But, as Ford & Sterman (1998, p. 309) write, "While many methods to elicit information from experts have been developed, most assist in the early phases of modeling: problem articulation, boundary selection, identification of variables, and qualitative causal mapping. … The literature is comparatively silent, however, regarding methods to elicit the information required to estimate the parameters, initial conditions, and behavior relationships that must be specified precisely in formal modeling."

18. This approach conforms to all seven of Breuker's (1987, summarized in Shadbolt & Burton, 1995) KADS (Knowledge Acquisition and Domain Structuring) principles for the elicitation of knowledge and construction of a system, as detailed in Nirenburg, McShane, & Beale (2010b).

19. For more on influence diagrams, see Howard & Matheson (2005). For an example of their use in another medical domain, see Lucas (1996).

20. For other issues related to reducing the complexity of knowledge acquisition of influence diagrams see Bielza et al. (2010).

21. This material draws from McShane, Nirenburg, & Jarrell (2013).

22. This observation was first made in Paul Meehl's (1996 [1954]) highly influential work that compared statistical predictions to clinical judgments and found the former to consistently outperform the latter. A recent review of Meehl's work (Grove & Lloyd, 2006) concludes that his findings have stood the test of time.

23. We call these MRs rather than TMRs because they need not derive from textual (T) inputs.

Chapter 9

1. A useful comparison can be made with the practices accepted in the *Advances in Cognitive Systems* community (cogsys.org). In its publication guidelines, the community recognizes that formal evaluations are not realistic at every stage of every program of work. Accordingly, the guidelines do not require that all conference and journal papers include formal evaluations. Instead, they require that papers adhere to other quite specific standards, including defining an important problem related to human-level intelligence or cognition, specifying theoretical tenets, making explicit claims, and supporting those claims—be it by argumentation, demonstration, or evaluation.

2. Jones et al. (2012, pp. 83–84) argue that cognitive systems need to integrate *adaptivity*, *directability*, *explainability*, and *trustworthiness* and that "evaluation requirements should, as nearly as possible, not be achievable by a system unless it addresses all four dimensions." Although we appreciate the spirit of this position, we find it too rigid. For both practical and scientific reasons, developers need to be able to evaluate and report intermediate successes as well.

3. There are typically multiple correct and appropriate ways of rendering a given text in another language. No prefabricated gold standard against which translations are measured can cover this space of paraphrases. If a system's translation does not textually match the gold standard, it does not necessarily mean that it is unacceptable. This observation motivates the justified criticisms of BLEU (Papineni et al., 2002; Callison-Burch et al., 2006), a widely used metric for evaluating machine translation systems.

4. Although in this book we subsume *multiword expressions* under the broader rubric of *constructions* (see section 4.3), the paper we cite treated a subset of phenomena that were appropriately referred to as *multiword expressions*.

5. The full TMRs were available to evaluators and were consulted as needed.

6. The negation is taken care of by a modality frame available in the full TMR.

7. Randomly selecting among same-scoring semantic analyses is only one of many possible system settings. The analyzer could also be configured to return all candidate analyses that score within a threshold of the highest score.

8. The software system reported there was subsequently replaced by a different one that incorporates the two kinds of incrementality described in this book. However, since the evaluation targets the underlying algorithms and knowledge bases, which stayed the same across implementations, we would expect similar results from the current system.

9. Joachim Eibl, http://kdiff3.sourceforge.net

10. If the field of NLP had not turned away from the problem of computing meaning some twenty-five years ago, we can imagine that the computational linguistics community might have, by now, made good progress toward this goal. However, as it stands, problems that were already identified and partially addressed became sidelined over the years, with full computational solutions remaining elusive.

11. Language complexity has been addressed from various perspectives. For example, the book *Measuring Grammatical Complexity* (Newmeyer & Preston, 2014) focuses on complexity as it pertains to theoretical syntax.