Linguistics for the Age of AI

The title-level DOI for this work is:

doi:10.7551/mitpress/13618.001.0001

9

Measuring Progress

Measuring progress is an important aspect of developing models and systems. But for knowledge-based approaches, evaluation is not a simple add-on to system development. Instead, inventing useful, practical evaluation methodologies is best viewed as an ongoing research issue. And, as a research issue, we should expect the process to be marked by trial and error. However, the associated successes, failures, and lessons learned are as central to the program of work as are the models and implementations themselves.

The reason why new evaluation approaches and metrics are needed is because the ones that have been adopted as the standard within the world of empirical NLP are a poor fit. In the chapter entitled "Evaluation of NLP Systems," published in *The Handbook of Computational Linguistics and Natural Language Processing* (Clark, Fox, & Lappin, 2010), Resnik and Lin (2010) do not even touch on the evaluation of knowledge-based or theoretically oriented NLP systems, writing (italics ours):

> It must be noted that the design or application of an NLP system is sometimes connected with a broader scientific agenda; for example, cognitive modeling of human language acquisition or processing. In those cases, the value of a system resides partly in the attributes of the theory it instantiates, such as conciseness, coverage of observed data, and the ability to make falsifiable predictions. Although several chapters in this volume touch on scientific as well as practical goals (e.g., the chapters on computational morphology, unsupervised grammar acquisition, and computational semantics), *such scientific criteria have fallen out of mainstream computational linguistics almost entirely in recent years in favor of a focus on practical applications, and we will not consider them further here*. (p. 271)

However, work outside the mainstream is ongoing, and those pursuing it have to take on the evaluation challenge. Jerry Hobbs, in "Some Notes on Performance Evaluation for Natural Language Systems" (2004), makes the following observations: Evaluation through demonstration systems is not conclusive because it can "dazzle observers for the wrong reasons"; evaluation through deployed systems that use emergent technologies is also not conclusive because such systems can fail to be embraced for reasons unrelated to the promise of the

technology; component-level evaluations are needed, but the competence and performance of systems must be considered separately, because "a system that represents significant progress in competence may be a disaster in performance for some trivial reason"; and, since there tends to be little uniformity of goals, foci, coverage, and applications across knowledge-based systems, head-to-head comparisons are well-nigh impossible.

Apart from the difficulty of formulating useful, representative evaluation suites, another important issue is cost. In their historical overview of evaluation practices in NLP, Paroubek et al. (2007, p. 26) point out that in the 1980s "the issue of evaluation was controversial in the field. At that time, a majority of actors were not convinced that the benefits outweighed the cost." They proceed to describe the subsequent emphasis on formal evaluation as a "trend reversal" but, curiously, do not overtly link this effect to its cause.

It seems clear that the focus on evaluation only became possible because the field at large adopted a system-building methodology based on statistical machine learning approaches that allowed individual developers to use standardized, straightforward, and inexpensive evaluation regimens of a particular kind. So this trend reversal toward an emphasis on evaluation says much more about the history of mainstream NLP than about the cost-benefit analysis of formal evaluations in principle.

Our point is not that knowledge-based programs of R&D should be absolved of providing evidence of progress—certainly not![1] However, the measures of progress adopted must be appropriate to the approaches and systems they evaluate, genuinely useful, and not so demanding of resources that they overwhelm the overall program of R&D.[2] This is the spirit of our ongoing efforts to measure progress on NLU within the broader program of work on developing humanlike LEIAs.

9.1 Evaluation Options—and Why the Standard Ones Don't Fit

Whatever methods one uses to build NLP capabilities, the top-level choice in evaluation is whether to evaluate an end application or a component functionality. Many believe that end-system evaluation is the gold standard. And, in fact, the evaluation practices that have become the standard in mainstream, empirical NLP originally grew out of such applications as information retrieval and information extraction. However, end-system evaluation is not always ideal. For example, if NLU capabilities are incorporated into a more comprehensive system, such as a robotic assistant, then the entire system—not only the NLU portion of it—needs to be at an evaluable stage of development, which can take a long time. Moreover, all NLP-specific capabilities required by the end system must also be developed and integrated prior to evaluation. Yet another drawback of end-system evaluations is that they are unlikely to address all language phenomena treated by the system, meaning that the evaluations say *something* about the system's capabilities but far from *everything interesting and useful*. Finally, error attribution can be difficult in an end system that is comprised of many diverse parts.

An alternative evaluation option focuses on individual components. The empirical NLP community has developed the practice of creating tasks that foster field-wide competitions targeting one or another linguistic phenomenon. The tasks—which are described in extensive guidelines—are formulated by individuals representing the community at large. Often, those individuals also oversee the compilation and annotation of corpora to support the training and evaluation of the supervised machine learning systems that will compete on the task. Among the many phenomena that have been addressed by task descriptions are coreference resolution, named-entity recognition, case role identification, and word-sense disambiguation. The fruits of this considerable, and expensive, task-formulation effort are made available to the community for free—something that stimulates work on the topic, allows for head-to-head comparisons between systems, and avoids the replication of effort across research teams. In short, these resources are of great utility to those whose goals and methods align with them.

However, to properly understand the role of tasks and task-oriented resources in the field, one must acknowledge not only their benefits but also their limitations. The task descriptions typically contain extensive listings of rule-in/rule-out criteria (e.g., Chinchor, 1997; Hirschman & Chinchor, 1997). The ruled-in instances are called *markables* because they are what annotators will mark in a corpus. Instances that are ruled out (not marked) are considered outside of purview.

The task descriptions reflect rigorous analysis by linguists, who must consider not only linguistic complexity but also the expected capabilities of annotators (often, college students), the speed/cost of annotation, the need for high interannotator agreement, and the anticipated strengths and limitations of the machine learning methods that are expected to be brought to bear. In many cases, the task description specifies that systems participating in an evaluation competition will be provided with annotated corpora not only for the training stage but also for the evaluation stage, which significantly distances the task from the full, real-world problem. And the more difficult instances of linguistic phenomena are usually excluded from purview because they pose problems for annotators and system developers alike. In sum, using the word *task* to describe such enterprises is quite appropriate; so is using such tasks to compare results obtained by different machine learning methods. It is important, however, to thoughtfully interpret—having read the task specifications—what the scores on associated evaluations mean. After all, 90% precision on a task does not mean 90% precision on automatically processing all examples representing the given linguistic phenomenon.

The reason for detailing the nature of mainstream NLP tasks was to make the following point. When we evaluate a LEIA's domain-neutral NLU capabilities using stages 1–5 of processing (*not* Situational Reasoning) over unrestricted corpora, our evaluation suites are no less idiosyncratic. They, too, cover only a subset of eventualities, and for the same reason—the state of the art is too young for any of us to do well on the very hardest of language inputs, and we all need credit for midstream accomplishments. There are, however,

significant differences between our pre-situational (stages 1–5) evaluations and mainstream NLP tasks.

1. Whereas mainstream tasks are formulated by community-wide representatives, we need to formulate our own. Community-level task formulation has three advantages that we do not share:

    a. The community gives its stamp of approval regarding task content and design, absolving individual developers of having to justify it.

    b. The task description exists independently and can be pointed to, without further discussion, by individual developers reporting their work, which facilitates the all-important publication of results.

    c. The community takes on the cost of preparing the task and all associated resources, so there is little to no cost to individual developers.

2. Mainstream tasks involve manually preselecting markables before systems are run, effectively making the difficult examples go away. We, by contrast, expose LEIAs to all examples but design them to operate with self-awareness. Just as people can judge how well they have understood a language input, so, too, must LEIAs. Relying on a model of metacognitive introspection that uses simpler-first principles (see section 2.6), LEIAs can automatically select the inputs that they believe they can treat competently. It is these inputs that are included in our evaluation runs. Requiring LEIAs to treat every example would be tantamount to requiring mainstream NLP tasks—and the associated annotation efforts—to treat every example as a markable. The field overall, no matter the paradigm (be it statistical NLP or knowledge-based NLU), is just too young for a *treat everything* requirement to be anything but futile. Recalling the discussion in section 1.6.3 (which juxtaposes NLU and NLP), we rightfully cheer for every individual behavior demonstrated by robots, not expecting them to be fully humanlike today. We need to shift the collective mindset accordingly when it comes to processing natural language.

    Note: We must reiterate that the evaluation setups we are talking about treat NLU outside the full cognitive architecture, applying only those knowledge bases and processors that cover the open domain (i.e., those belonging to stages 1–5 of LEIA operation). The above juxtaposition with mainstream NLP tasks is meant to stress that evaluating pre-situational, open-domain NLU by LEIAs is very different from evaluating full NLU within an end application. Within end applications, LEIAs have to treat every input but can take advantage of (a) specialized domain knowledge, (b) Situational Reasoning (stage 6), and (c) the ability to decide how precise and confident an analysis must be to render it actionable.

3. Whereas NLP task suites include a manually annotated gold standard against which to evaluate system performance, most of our evaluation experiments—namely, those

requiring TMR generation—have involved checking the system's output after it was produced. The reason why is best understood by considering the alternatives.

a. If people were asked to manually create gold standard TMRs on the basis of the ontology alone (i.e., without the lexicon), this gold standard would be suboptimal for evaluation because of the possibility of ontological paraphrase. That is, the system might generate a perfectly acceptable TMR that did not happen to match the particular paraphrase listed in the gold standard. This is similar to the problem of accounting for linguistic paraphrase when evaluating machine translation systems.[3]

b. If people were told to use the lexicon and ontology together to create gold standard TMRs, then they would be carrying out a very inefficient replication of the automatic process. It is for good reason that the time and cost of annotation has always been at the center of attention in statistical NLP. We cannot collectively afford to spend unbounded resources on evaluation—particularly if they would be as ill-used as under this scenario.

4. Whereas mainstream task formulation involves teams of people carrying out each aspect of manual data preparation (with interannotator agreement being an important objective), we do not have comparable resources and so must find alternative solutions.

5. Whereas mainstream task-oriented evaluations are black box and geared at generating numerical results to facilitate comparisons across systems, ours are glass box and only partially numerical. Our emphasis is on understanding the reasons for particular outcomes, which is necessary to assess the quality of our models, to determine the success of the model-to-system transition, and to chart directions for future development.

The following are among the approaches to evaluation that knowledge-based efforts can adopt.

1. *Carrying out evaluations that target specific phenomena within small domains*. This has been done, for example, in the work of James Allen and collaborators (e.g., Allen et al., 2006, 2007; Ferguson & Allen, 1998).

2. *Wearing two hats: scientific and technological*. In their role as scientists, developers carry out cognitively inspired, rigorous descriptive work, but in their role as technologists, they select simplified subsets of phenomena for use in application systems that are evaluated using the traditional NLP approach. This appears to be the choice of the dialog specialist David Traum (compare Traum, 1994, for scientific work with Nouri, Artstein, Leuski, & Traum, 2011, for application-oriented work).

3. *Building theories but not applying them in computational systems*. This approach—which is typical, for example, of computational formal semanticists—has been criticized on the grounds that NLP must involve actual computation (see, e.g., Wilks, 2011). However, its motivation lies in the promise of contributing to future system building.

4. *Pursuing hybrid evaluations*. Hybrid evaluations combine aspects of the above approaches, which we find most appropriate for evaluating NLU by LEIAs.

The sections to follow describe our team's experience with evaluation. It includes five component-level (i.e., microtheory-oriented) evaluation experiments (section 9.2) and two holistic ones (section 9.3). We describe the experiments in some detail because we believe that our experience will be of use to others undertaking evaluation as part of R&D in NLU. All the evaluations we describe were carried out on unrestricted corpora, with varying rules of the game that we will specify for each evaluation. In all cases, the experiments validated that our system worked essentially as expected. But the real utility of the experiments lay in the lessons learned—lessons that would have been unavailable had we not actually implemented our models, tested them on real inputs, and observed where they succeeded and failed. Introspection, no matter how informed by experience, just does not predict all the ways people actually use language.

The most important lesson learned was that, with higher-than-expected frequency, the interpretation of an input can seem to work out well (i.e., receive a high-confidence score) yet be incorrect. For example, an agent cannot be expected to guess that *kick the bucket* or *hit the deck* have idiomatic meanings if those meanings are not recorded in the lexicon, since it is entirely possible to strike a bucket with one's foot and slap a deck with one's hand. However, once such meanings are recorded, agents can include the idiomatic readings along with the direct ones in the analysis space. Although adding a lexical sense or two would be a simple fix for many attested errors, what is needed is a much more comprehensive computational-semantic lexicon than is currently available. Building such a lexicon is an entirely doable task, but, in the current climate, it is unlikely to be undertaken at a large scale because the vast majority of resources for human knowledge acquisition fieldwide are being devoted to corpus annotation. So the "seems right but is wrong" challenge to NLU systems operating in the open domain will remain for the foreseeable future.

9.2 Five Component-Level Evaluation Experiments

Over the past several years we formally evaluated our microtheories for five linguistic phenomena: nominal compounding; multiword expressions;[4] lexical disambiguation and the establishment of the semantic dependency structure; difficult referring expressions; and verb phrase ellipsis. Each of these evaluation experiments played a minor part in a published report whose main contribution was the microtheory itself—that is, the description of a model, grounded in a theory, that advances the fields of linguistics and computational cognitive modeling. However, it was important to include the description of an evaluation experiment to show that the microtheories were actually computational. The challenge in each case was to carve out a part of the microtheory that could be teased apart relatively cleanly from all the other interdependent microtheories required for comprehensive NLU.

This section presents a sketch of each of those evaluations. We do not repeat the numerical results for three reasons: they are available in the original papers; their precise interpretation requires a level of detail that we are not presenting here; and we don't believe that a theoretically oriented book, which should have a reasonably long shelf life, should include necessarily fleeting progress reports.

It is worth noting that all of these evaluation setups were deemed reasonable by at least those members of the community who served as reviewers for the respective published papers. Our hope is that these summaries highlight the unifying threads across experiments without losing the aspects of those original reports that made them convincing.

9.2.1 Nominal Compounding

Our evaluation of the microtheory of nominal compounding (McShane et al., 2014) focused on lexical and ontological constructions that both disambiguate the component nouns and establish the semantic relationship between their interpretations. That is, if two nouns can be interpreted using the expectations encoded in a recorded construction, then it is likely that they should be interpreted using that construction. These constructions were described in section 6.3.1.

For example, the nouns in the compound *bass fishing* are ambiguous: bass [BASS-FISH, STRING-BASED-INSTRUMENT], fishing [FISHING-EVENT, SEEK]. Combining these meanings leads to four interpretations:

- Carrying out the sport/job of fishing in an attempt to catch a type of fish called a *bass*;
- Carrying out the sport/job of fishing in an attempt to catch a stringed musical instrument called a *bass*;
- Seeking (looking for) a type of fish called a *bass*; or
- Seeking (looking for) a stringed musical instrument called a *bass*.

However, only one of these interpretations, the first, matches a recorded NN construction, namely, FISH + *fishing* → FISHING-EVENT (THEME FISH). By analyzing *bass fishing* according to this construction, the system simultaneously selects a meaning of *bass*, a meaning of *fishing*, and the relationship between them. The existence of this construction asserts a preference for this interpretation as the default. We must emphasize that this is still only a tentative, default interpretation that may be discarded when the analysis of the nominal compound is incorporated into the clause-level semantic dependency structure. In the reported evaluation, we assessed how often this default interpretation was correct.

This evaluation did not address all aspects of the microtheory of nominal compounding, such as processing compounds containing three or more nouns or compounds in which one or both of the words are unknown and need to be learned on the fly. This would have invoked not only new-word learning capabilities but also all the aspects of analysis contributing to it, such as clause-level lexical disambiguation and coreference resolution.

The corpus used for evaluation was the *Wall Street Journal* (1987; hereafter, WSJ). String-search methods identified sentences of potential interest, and those candidates remained in the evaluation corpus if they met all of the following criteria:

1. The sentence could be analyzed, with no technical failures, by the CoreNLP preprocessor, the CoreNLP syntactic dependency parser, and the LEIA's semantic analyzer. If there was a failure, then the given sentence was automatically excluded from purview. It is not feasible to turn the evaluation of a particular microtheory into the evaluation of every system component—particularly those, like CoreNLP, that we import.
2. The NN string was recognized by the parser as a compound, it contained exactly two nouns, and neither of those was a proper noun or an unknown word.
3. The semantic analyses of both the NN and the verb that selected it as an argument were headed by an ontological concept rather than a modality frame, a call to a procedural semantic routine, or a pointer to a reified structure. This made the manual inspection of the system's results reasonably fast and straightforward.
4. The NN served as an argument of the main verb of the clause, which permits clause-level disambiguation using selectional constraints. If the NN was, for example, located in a parenthetical expression or used as an adjunct, then disambiguation would rely much more heavily on reference resolution and extrasentential context.

This pruning of candidate contexts was carried out automatically, with supplementary manual inspection to weed out processing errors (e.g., not recognizing that a compound contained three, not two, nouns). After this pruning, 72% of the examples initially extracted were deemed within purview of the evaluation, resulting in 935 examples.

The manual checking of the system's results was carried out by a graduate student under the supervision of a senior developer. The manual vetting involved reading the portion of the TMR(s) that represented the meaning of the compound and determining whether it was correct in the context.

The evaluation results overall were positive: the system returned the appropriate decision when it could be expected to do so. What is most interesting is what the system got wrong and why. There were three main sources of errors—lexical idiosyncrasy, polysemy/ambiguity, and metaphorical usage—which we describe in turn.

Lexical idiosyncrasy. Most mistakes involved lexically idiosyncratic compounds—that is, ones whose meanings need to be explicitly recorded in the lexicon rather than dynamically computed using standard expectations. For example (in plain English rather than the ontological metalanguage):

1. *Talk program* was incorrectly analyzed as a social event whose purpose was either conversation or lecturing—as might be plausible, for example, as an activity for nursing home residents to keep them socially active. The intended meaning, however, was a radio or TV program that involves talking rather than, say, music or drama.

2. *College education* and *public education* were incorrectly analyzed as teaching about college and society, respectively, using the construction that would have been correct for *science education* or *history education*.
3. *Pilot program* was analyzed as a social event that benefits airplane pilots, which is actually plausible but is not the meaning (*feasibility study*) that was intended in the examples.
4. *Home life* was analyzed as the length of time that a dwelling could be used, employing a construction intended to cover compounds like *battery life* and *chainsaw life*.

In some cases, these errors pointed to the need to further constrain the semantics of the variables in the construction that was selected. However, more often, the compounds simply needed to be recorded as constructions in the lexicon along with their not-entirely-predictable meanings.

Polysemy/ambiguity. In some cases, a compound allowed for multiple interpretations, even though people might zero in on a single one due to its frequency, their personal experience, or the discourse context. When the system recognized ambiguities, it generated multiple candidate interpretations. To cite just a few examples:

1. *Basketball program* was analyzed as a program of activities dedicated either to basketballs as objects (maybe they were being donated) or to the game of basketball.
2. *Oil spill* was analyzed as the spilling of either industrial oil or cooking oil.
3. *Ship fleet* was analyzed as a set of sailing ships or spaceships.

A curious ambiguity-related error was the analysis of *body part* as a part of a human, since that compound was recorded explicitly in the lexicon during our work on the MVP application. In a particular corpus example, however, it referred to a car part.

As mentioned earlier, the meaning of compounds must be incorporated into the meaning of the discourse overall, and this is part of our full microtheory of compounding. However, in order to keep this experiment as simple and focused as possible, we put the system at an unfair disadvantage. We forced it to accept the default NN interpretation that was generated using a construction without allowing it to reason further about the context; yet we penalized it if that default interpretation was incorrect! This is a good example of the trade-offs we must accept when, for purposes of evaluation, we extract particular linguistic phenomena from the highly complex, multistage process of NLU.

Metaphorical usage. Metaphorical uses of NNs are quite common. For example, in (9.1) both *rabbit holes* and *storm clouds* are used metaphorically.

(9.1) He also alerts investors to key financial *rabbit holes* such as accounts receivable and inventories, noting that sharply rising amounts here could signal *storm clouds* ahead. (WSJ)

In some cases, automatically detecting metaphorical usage is straightforward, as when the NN is preceded by the modifier *proverbial*.

(9.2) "They have taken the proverbial *atom bomb* to swat the fly," says Vivian Eveloff, a government issues manager at Monsanto Co. (WSJ)

In other cases, it can be difficult to detect that something other than the direct meaning is intended.

In addition to non-compositionality and residual ambiguity, our work on NN compounding has revealed other challenges. For example, certain classes of compounds are very difficult to semantically analyze, even wearing our finest linguistic hats. A star example involves the headword *scene*, used in compounds such as *labor scene*, *drug scene*, and *jazz scene*. The meanings of these compounds can only be adequately described using full ontological scripts—a different script for each kind of scene. Anything less, such as describing the word *scene* using an underspecified concept like SCRIPT-INDICATOR, would just be passing the buck.

Even if NNs are not as semantically loaded as *scene* compounds, many more than one might imagine are not fully compositional and, therefore, must be recorded as fixed expressions. In fact, most of the NNs that we recorded as headwords in the lexicon prior to the evaluation study were analyzed correctly, which suggests that our lexicalization criteria are appropriate. Of course, occasionally we encountered an unforeseen point of ambiguity, as in the case of *body part*, referred to earlier.

To summarize the NN compounding experiment:

- It validated the content and utility of the portion of the microtheory tested.
- The system worked as expected; that is, it faithfully implemented the model.
- The lexicon needs to be bigger: there is no way around the fact that language is in large part not semantically compositional.
- The experimental setup did not address the need for contextual disambiguation of nominal compounds.
- It can be difficult to automatically detect certain kinds of mistakes when the wrong interpretation seems to work out fine, as in the case of NNs being used metaphorically.

Spoiler alert: This list of experimental outcomes will largely be the same for the rest of the experiments we describe here.

9.2.2 Multiword Expressions

As explained in section 4.3, there is no single definition of *multiword expression* (MWE). For the evaluation reported in McShane, Nirenburg, and Beale (2015), we defined *MWEs of interest* as those lexical senses whose syn-struc zones included one or more specific words

that were not prepositions or particles. For example, *cast-v3* (*X {cast} a spell on/over Y*) requires the direct object to be the word *spell*; similarly, *in-prep15* (*X be in surgery*) requires the object of the preposition to be the word *surgery*. Neither the inventory of MWEs covered in the lexicon nor the lexicon entries themselves were modified before evaluation: all evaluated MWEs were recorded during regular lexical acquisition over prior decades.

The evaluation worked as follows. The system automatically identified 382 MWEs of interest in our lexicon and then used a string-based (nonsemantic) method to search the *Wall Street Journal* corpus of 1987 for sentences that might contain them—"might" because the search method was underconstrained and cast a wide net. The requirements of that search were that all lexically specified roots in the MWE occur within six tokens of the headword. For example, to detect candidate examples of the MWE *something {go} wrong with X*, the word *something* had to be attested within six tokens preceding *go/went/goes*, and the words *wrong* and *with* had to be attested within six tokens following *go/went/goes*. This filtering yielded a corpus of 182,530 sentences, which included potential matches for 286 of our 382 target MWEs. We then selected the first 25 candidate hits per MWE, yielding a more manageable set of 2,001 sentences, which were syntactically parsed. If the syntactic parse of a sentence did not correspond to the syntactic requirements of its target MWE—that is, if the actual dependencies returned by CoreNLP did not match the expected dependencies recorded in our lexicon—the sentence was excluded. (Recall that the initial candidate extraction method was quite imprecise—we did not expect it to return exclusively sentences containing MWEs.) This pruning resulted in 804 sentences that syntactically matched 81 of our target MWEs. We then randomly selected a maximum of 2 sentences per target MWE, resulting in an evaluation corpus of an appropriate size: 136 sentences.

These 136 sentences were semantically analyzed in the usual way. The analyzer was free to select any lexical sense for each word of input, either using or not using MWE senses. To put the lexical disambiguation challenge in perspective, consider the following:

- The average sentence length in the evaluation corpus was 22.3 words.
- The average number of word senses for the headword of an MWE was 23.7. This number is so high because verbs such as *take* and *make* have over 50 senses apiece due to the combination of productive meanings and light-verb usages (e.g., *take a bath*, *take a nap*, *take sides*).
- The average number of word senses for each unique root in the corpus was 4.

To summarize, the system was tasked with resolving the syntactic and semantic ambiguities in these inputs using an approximately 30,000-sense lexicon that was not tuned to any particular domain. One developer manually inspected the system's results and another carried out targeted double-checking and selective error attribution.

Since the TMRs for long sentences can run to several pages, we used a TMR-simplification program to automatically extract the minimal TMR constituents covered by the candidate

MWE. For example, in (9.3) and (9.4), the listed TMR excerpts were sufficient to determine that the MWEs (whose key elements are in italics) were treated correctly.[5]

(9.3) The company previously didn't *place* much *emphasis on* the development of prescription drugs and relied heavily on its workhorse, Maalox. (WSJ)

EMPHASIZE-1[6]
AGENT FOR-PROFIT-CORPORATION-1
THEME DEVELOP-1

(9.4) "I'm sure nuclear power is good and safe, but it's impossible in the Soviet bloc," says Andrzej Wierusz, a nuclear-reactor designer who *lost* his *job* and was briefly jailed after the martial-law crackdown of 1981. (WSJ)

ASPECT-1
SCOPE WORK-ACTIVITY-1
PHASE end

WORK-ACTIVITY-1
AGENT HUMAN-1

The questions posited in the evaluation were:

- Did the system correctly identify sentences in which an MWE was used?
- Did it correctly compute the meaning of the MWE portion of those sentences?

Note that the latter does not require correctly disambiguating all words filling variable slots in MWEs, since that would complicate the evaluation tenfold, forcing it to cover not only lexical disambiguation overall but also coreference resolution.

In most cases, these evaluation criteria meant that the EVENT head of the TMR frame representing the MWE's meaning needed to be selected correctly—such as EMPHASIZE in (9.3). But in some cases, multiple elements contribute to the core meaning of an MWE, so all of them needed to be correct. For example, in (9.4) the combination of the ASPECT frame—with its "PHASE end" property—and the WORK-ACTIVITY frame represents the core meaning of the MWE.

The decision about correctness was binary. If the needed TMR head was (or heads were) correct, then the MWE interpretation was judged correct; if not, the MWE interpretation was judged incorrect.

In many cases, the system correctly analyzed more than what was minimally needed for this evaluation. For example, in (9.3) it correctly disambiguated the fillers of the AGENT and THEME case roles. It selected FOR-PROFIT-CORPORATION as the analysis of the ambiguous word *company* (which can also refer to a set of people), and it selected DEVELOP-1 as the analysis of *development* (which can also refer to a novel event or a residential area). However, to reiterate, we did not require that case role fillers be correctly disambiguated

in order to mark an MWE interpretation as correct because this can require much more than clause-level heuristics. For example, in (9.5), the MWE analysis was correct: *to look forward to X* means (roughly) to *want* the event or state of affairs X to occur, which is represented in the TMR by the highest value of volitive modality scoping over X.

(9.5) We *look forward to* the result.

MODALITY-1	
SCOPE	ANY-NUMBER-1
VALUE	1
ATTRIBUTED-TO	SET-1
TYPE	volitive

However, the filler of one of the slots in this TMR—the SCOPE of the modality—is probably not correct. This TMR says that what was looked forward to was some number (ANY-NUMBER), which is one sense of *result*, whereas what is probably being looked forward to is some state of affairs—another sense of *result*. However, in all fairness, the system *could* be correct since the sentence could be uttered in a math class by students waiting for their resident genius to solve a problem. Lacking extra-clausal heuristic evidence, the system arrived at comparable scores for both analyses and randomly selected between them.[7]

Examples (9.6) and (9.7) offer further insights into why we did not fold into the evaluation the disambiguation of case role fillers. All four salient case roles in these examples of the MWE *X {pose} problem for Y* were analyzed incorrectly, even though analysis of the MWE was correct.

(9.6) The changing image did however *pose* a *problem* for the West. (WSJ)

(9.7) But John McGinty, an analyst with First Boston Corp., said he believed dissolution of the venture won't *pose* any *problem* for Deere. (WSJ)

Two of the errors—the analyses of *the West* and *Deere*—were due to the mishandling of proper names (something handled by the CoreNLP tool set, whose preprocessing results we import). One error—the analysis of *the changing image*—could not be correctly disambiguated using the sentence-level context provided by our examples: that is, *image* can be a pictorial representation or an abstract conceptualization. And the final error—the analysis of *dissolution of the venture*—results from a failure to simultaneously recognize the metaphorical usage of *dissolve* and select the correct sense of the polysemous noun *venture*. These examples underscore just how many different factors contribute to making NLU as difficult as it is.

In some cases, the system did not select the MWE sense of a lexical item that it should have preferred. Instead, it analyzed the input compositionally. The reasons were not always apparent, apart from the fact that the scoring bonus for MWE analyses over compositional ones is relatively minor. Clearly, other preferences in the analysis of the sentence overall had a deciding role.

To reiterate a point made earlier, the system had to select from an average of 23.7 word senses for each MWE head, each having their own inventories of expected syntactic and semantic constraints, which competed to be used in the analysis of each input. So, although our approach to NLU and the system implementation are as transparent as they can be, the effects of combinatorial complexity cannot always be untangled.

Many of the errors in processing MWEs can be obviated through additional knowledge acquisition: namely, by acquiring more MWEs, by adding more senses to existing MWEs, and by more precisely specifying the rule-in/rule-out constraints on MWEs. This was discussed in section 4.3.5.

As promised, we can summarize the results of this experiment using the same points as for the nominal compounding experiment:

- The experiment validated the content and utility of the portion of the microtheory tested.
- The system worked as expected.
- The lexicon needs to be bigger and some of its entries need to be more precisely specified.
- Some problems, such as residual ambiguity, need to be resolved by methods that were not invoked for the experiment.
- It is difficult to automatically detect certain kinds of mistakes when the wrong interpretation seems to work fine, as in the case of metaphorical usage.

9.2.3 Lexical Disambiguation and the Establishment of the Semantic Dependency Structure

The experiment reported in McShane, Nirenburg, and Beale (2016) focused on the system's ability to carry out lexical disambiguation and establish the semantic dependency structure.[8] As always, we attempted to give the system a fair opportunity to demonstrate its capabilities while neither overwhelming it with complexity nor reducing the endeavor to a toy exercise. The system was required to

1. disambiguate head verbs: that is, specify the EVENT needed to express their meaning as used in the context; and
2. establish which case roles were needed to link that EVENT to its semantic dependents.

We did not evaluate the disambiguation of the fillers of case role slots for the same reason as was described earlier: this often requires coreference resolution and/or other aspects of discourse analysis that would have made the evaluation criteria impossibly complicated.

The evaluation corpus included four Sherlock Holmes stories: "A Scandal in Bohemia," "The Red-Headed League," "A Case of Identity," and "The Boscombe Valley Mystery"

(hereafter referred to collectively as *S-Holmes*). We selected these because they are freely available from Project Gutenberg (EBook #1661) and, to our knowledge, nobody has recorded linguistic annotations of these works, so there can be no question that the system operated on unenhanced input.

We first selected an inventory of verbs of interest from our lexicon, all of which had the following two properties: (a) they had at least two senses, so that there would be a disambiguation challenge, and (b) those senses included syntactic and/or semantic constraints that allowed for their disambiguation. Ideally, all lexical senses would include such disambiguating constraints, but this is not always possible. A frequent confounding case involves pairs of physical and metaphorical senses that take the same kinds of arguments. For example, if person A attacks person B, A might be physically assaulting or criticizing B, something that can only be determined using additional knowledge about the context.

The system automatically selected, and then semantically analyzed, 200 sentences containing verbs that corresponded to the selection criteria. We then manually checked the correctness of the resulting TMRs. One developer carried out this work with selective contributions from another. The evaluation involved not only identifying errors but also attempting to trace them back to their source so that they could be fixed to improve future system functioning. We did not amend the lexicon or ontology in any way to prepare for this evaluation.

The experimental setup included challenges of a type that are often filtered out of mainstream NLP evaluation suites. For instance, some examples did not contain sufficient information to be properly disambiguated, as by having semantically underspecified pronouns fill key case roles; other examples reflected what might be considered nonnormative grammar. However, considering the importance of automatically processing nonstandard language genres (texting, email, blogs), we felt it appropriate to make the system responsible for all encountered phenomena.

The two main sources of errors, beyond singletons that are of interest only to developers, were lexical lacunae and insufficiencies of the experimental design, which we describe in turn.

Lexical lacunae. Most disambiguation errors resulted from the absence of the needed lexical sense in the lexicon. Often, the missing sense was part of an idiomatic construction that had not yet been acquired, such as *draw the blinds* in (9.8).

(9.8) The *drawn* blinds and the smokeless chimneys, however, gave it a stricken look. (S-Holmes)

In other cases, the needed semantic representation (sem-struc) was available in the lexicon but it was not associated with the needed syntactic realization (syn-struc). For example, for the system to correctly process (9.9), the lexicon must permit *announce* to take a direct object. However, the sense available in the lexicon—which did semantically describe the needed meaning of *announce*—required a clausal complement.

(9.9) She became restive, insisted upon her rights, and finally *announced* her positive intention of going to a certain ball. (S-Holmes)

A trickier type of lexical lacuna involves grammatical constructions that are not sufficiently canonical (at least in modern-day English) to be recorded in the lexicon. For example, the verb *pronounce* in (9.10) is used in the nonstandard construction *X pronounces Y as Z*.

(9.10) "I found the ash of a cigar, which my special knowledge of tobacco ashes enables me to *pronounce* as an Indian cigar." (S-Holmes)

Such sentences are best treated as unexpected input, to be handled by the recovery procedures described in section 3.2.4.

Insufficiencies of experimental design. Since we did not invoke the coreference resolution engine for this experiment, we should have excluded examples containing the most underspecified pronominal case role fillers: *it*, *they*, *that*, and *this*. (By contrast, personal pronouns that most often refer to people—such as *he*, *she*, *you*, and *we*—are not as problematic.) An argument like *it* in (9.11) is of little help for clause-level disambiguation.

(9.11) I walked round it and *examined it* closely from every point of view, but without noting anything else of interest. (S-Holmes)

For this example, the system selected the abstract event ANALYZE, which expects an ABSTRACT-OBJECT as its THEME. It should have selected the physical event VOLUNTARY-VISUAL-EVENT, which expects a PHYSICAL-OBJECT as the THEME. Of course, if this experiment had included coreference and multiclause processing, then the direct object of *examined* would corefer with the previous instance of *it*, which must refer to a physical object since it can be walked around.

Were the results of, and lessons learned from, this experiment the same as for the previous ones? Indeed, they were.

- The experiment validated the content and utility of the portion of the microtheory tested.
- The system worked as expected.
- The lexicon needs to be bigger.
- Some problems, such as residual ambiguity, need to be resolved by methods that were not invoked for the experiment.
- It is difficult to automatically detect certain kinds of mistakes when the wrong interpretation seems to work fine (as in the case of metaphorical usage).

One additional note deserves mention. Since most errors were attributable to missing or insufficiently precise verbal senses, and since we used the verbs in our lexicon to

guide example selection, we could have avoided most mistakes by using a different experimental setup. That is, before the evaluation we could have optimized the inventory of lexical senses for each selected verb, particularly by boosting the inventory of recorded multiword expressions. This would have required some, but not a prohibitive amount of, acquisition time. It would likely have substantially decreased the error rate, and it would likely have better highlighted the system's ability to manipulate competing syntactic and semantic constraints during disambiguation. However, an experiment of this profile would have less realistically conveyed the current state of our lexicon since we would not have done that enhancement for all of its verbs, not to mention all of the verbs in English. It would be hard to argue that either of these task formulations is superior to the other given that both would confirm the core capability of lexical disambiguation that was being addressed.

9.2.4 Difficult Referring Expressions

McShane and Babkin (2016a) describe the treatment and evaluation of two classes of referring expressions that have proven particularly resistant to statistical methods: broad referring expressions (e.g., pronominal *this*, *that*, and *it*; see section 5.3) and third-person personal pronouns (see section 5.2).

Broad referring expressions are difficult not only because they can refer to spans of text of any length (i.e., one or more propositions) but also because they can refer to simple noun phrases, and the system does not know a priori which kind of sponsor it is looking for. Third-person personal pronouns, for their part, are difficult because semantic and/or pragmatic knowledge is often required to identify their coreferents. We prepared the system to treat difficult referring expressions by defining lexico-syntactic constructions that predicted the coreference decisions. These constructions do not cover a large proportion of instances in a corpus (i.e., they have low recall), but they have proven useful for what they do cover.

For the evaluation, the system had to (a) automatically detect, in an unrestricted corpus, which instances of difficult referring expressions matched a recorded construction and then (b) establish the coreference link predicted by that construction. This evaluation is more difficult to summarize than others because each construction was evaluated individually. That is why select evaluation results were reported in the sections that introduced the microtheories themselves (sections 5.2.2 and 5.3).

For the development and evaluation portions of this experiment, we used different portions of the English Gigaword corpus (Graff & Cieri, 2003; hereafter, Gigaword). For the first time, we compiled a gold standard against which the system would be evaluated. This involved two steps. First, the system identified the examples it believed it could treat confidently (since they matched recorded constructions). Then two graduate students and one undergraduate student annotated those examples according to the following instructions:

[NE]	If the selected entity is not actually a referring expression (e.g., pleonastic *it*), type [NE] before the example.
[ ]	If there is a single perfect or near-perfect antecedent, surround it with brackets.
[Mult]	If there is more than one possible antecedent, type [Mult] before the example and use multiple sets of brackets to indicate the options.
[Close]	If an available text string is close to the needed antecedent but not a perfect match, type [Close] before the example and use brackets to show the best available antecedent.
[Impossible]	If no text string captures the meaning of the antecedent, type [Impossible] before the example.
[Prob]	If there is some other problem with the context (e.g., it is unintelligible) type [Prob] before the example.

Annotators were shown a few worked examples but given no further instructions. This contrasts with the mainstream NLP annotation efforts that involve extensive guidelines that are painstakingly compiled by developers and then memorized by annotators.

When the annotation results were in, senior developers manually reviewed them (with the help of the program KDiff3[9]) and selected which ones to include in the gold standard. Often we considered more than one result correct. Occasionally, we added an additional correct answer that was not provided by the annotators. As expected, there was a considerable level of interannotator disagreement, but most of those differences were inconsequential. For example, different annotators could include or exclude a punctuation mark, include or exclude a relative clause attached to an NP, include or exclude the label [Close], or select different members of a coreference chain as the antecedent. We did not measure interannotator agreement because any useful measure would have required a well-developed approach to classifying important versus inconsequential annotation decisions—something that we did not consider worth the effort. To evaluate the system, we semiautomatically (again, with the help of KDiff3) compared the system's answers to the gold standard, calculated precision, and carried out error analysis toward the goal of system improvement.

As with previous experiments, this one validated the content and utility of the portion of the microtheory tested, and the system worked as expected. It pointed to the need for additional knowledge engineering on the constructions themselves, particularly on specifying rule-out conditions. This evaluation differed from previous ones in that we first created a gold standard and then tested system results against it. That process was more expensive and time-consuming than our previous approaches to vetting system outputs, but it was not prohibitively heavy because the decision-making about the coreference relations was relatively straightforward. However, as we will see in the next section, applying the same gold standard–first methodology to the task of VP ellipsis was a different story entirely.

9.2.5 Verb Phrase Ellipsis

To date, we have carried out two evaluations of different iterations of our model for VP ellipsis. The first, reported in McShane and Babkin (2016b), treated only elided VPs, whereas the second, reported in McShane and Beale (2020), treated both elided and

overt-anaphoric VPs—the latter realized as *do it*, *do this*, *do that*, and *do so*. We use the former experiment for illustration because it involved a more formal evaluation setup and, therefore, offers more discussion points for this chapter on evaluation.

The 2016 system—called ViPER (Verb Phrase Ellipsis Resolver)—had to

1. identify instances of VP ellipsis in an unconstrained corpus (Gigaword);
2. determine which instances it could treat using its repertoire of resolution strategies; and
3. identify the text string that served as the sponsor.

As our original report explains, this definition of *resolution* is partial in that it does not account for the important semantic decisions that we describe in section 5.5. However, the ability of this module to detect which contexts can be treated by available resolution strategies and to point out the sponsor counts as a significant contribution to the very demanding challenge of full VP ellipsis resolution.

The aspect of this experiment that is most salient to this chapter involves the repercussions of our decision to create a gold standard first, by annotating examples in the way that is traditional for mainstream (machine learning–oriented) NLP tasks. We anticipated a lot of eventualities and incorporated them into the annotation instructions. For example:

- Some of the examples that the system selected to treat might not actually be elliptical.
- The sponsor might be outside the provided context.
- There might be no precisely correct sponsor in the linguistic context at all.
- There might be multiple reasonable sponsor selections.

A few examples will serve to illustrate tricky cases, with their complexities indicated in square brackets.

(9.12) [Either of the previous mentions in the chain of coreference is a valid sponsor.]
However, Beijing still [rules the country with harsh authoritarian methods] in the provinces and will [continue to do so] for as long as it can __. (Gigaword)

(9.13) [The direct object could be included (which is more complete) or excluded (which sounds better).]
Nuclear power may [[give] NASA's long-range missions] the speed and range that combustion engines can not __, but research is sputtering for lack of funds. (Gigaword)

(9.14) [The first conjunct, 'go out and', may or may not be considered part of the sponsor.]
"We had to [go out and [play the game]] just like they did __." (Gigaword)

(9.15) [The actual sponsor is the noncontiguous 'pull off'; we did not allow for noncontiguous sponsors in order to avoid complexity, but this decision had some negative consequences.]
"I feel I can [[pull] that shot off]; that's just one of those I didn't __." (Gigaword)

(9.16) [The sponsor can, itself, be elided. Here, the actual resolution should be 'let them disrupt us'.]
"They can disrupt you if you [let them], and we didn't __." (Gigaword)

It would have taken a very detailed, difficult-to-master set of annotation rules to ensure that annotators were highly likely to make the same sponsor selection.

Given our lenient annotation conventions, for 81% of the examples in the evaluation suite (320 out of 393), all student annotators agreed on the sponsor, and that answer was considered correct (i.e., it was not vetted by senior developers). For the other 73 examples, senior developers had to decide which answer(s) qualified as correct. Then, in order to make the evaluation results as useful as possible, we created guidelines to judge ViPER's answers as correct, incorrect, or partially correct.

Correct required that the answer be exactly correct. *Incorrect* included three eventualities:

- Sentences that ViPER thought were elliptical but actually were not;
- Sentences whose sponsor was not in the provided context but ViPER pointed to a sponsor anyway; or
- Cases in which ViPER either did not identify the head of the sponsor correctly or got too many other things wrong (e.g., the inclusion or exclusion of verbs scoping over the sponsor head) to qualify for partial credit.

The second eventuality is actually the most interesting since it represents a case that would never make it into traditional evaluation tasks—that is, the case in which the answer is not available and the system is required to understand that. In traditional evaluation setups, examples that are deemed by task developers to be too difficult or impossible are excluded from the start.

Partial credit covered several eventualities, all of which involved correctly identifying the verbal head of the sponsor but making a mistake by including too many other elements (e.g., modal scopers) or excluding some necessary ones. Getting the sponsor head right is actually a big deal because it shows not only that the system can identify the sponsor clause but also that it understands that the example is, in principle, within its ability to treat.

It is important to note that many of the problems of string-level sponsor selection simply go away when the full NLU system is invoked, since actual VP resolution is done at the level of TMRs (semantic analyses), not words.

ViPER's methods for identifying treatable cases of VP ellipsis and identifying the sponsor worked well, as the evaluation numbers reported in the paper show. For reasons described earlier, the system itself chose what to treat and what not to treat, and we made no attempt to calculate its recall over the entire corpus. This would actually not have been trivial because our VP-detection process did not attempt full recall—ellipsis detection

being a difficult problem in its own right. Instead, our goal in developing detection methods was to compile a useful corpus with relatively few false positives.

In terms of system operation, this experiment yielded no surprises. However, we did learn to think thrice before undertaking any more annotation-first approaches to evaluating system operation for a problem as complex as VP ellipsis. At least in this case, the game was not worth the candle. We would have obtained the same information about the problem space and system operation if developers had reviewed system results without a precompiled gold standard. In fact, when it came time to do our next evaluation of VP ellipsis resolution (along with overt-anaphoric VP resolution), we did not create a gold standard first. Instead, developers vetted the results, and we called the process a *system-vetting experiment* rather than an evaluation (McShane & Beale, 2020). In fact, the latter approach was not only faster and cheaper but also more useful than the evaluation just described because we were not bound to one round of experimentation for reasons of cost/practicality. Instead, we iteratively developed and vetted the model and system in a way that best served our scientific and engineering goals.

9.3 Holistic Evaluations

The evaluation reported in McShane et al. (2019)—as well as a follow-up, unpublished evaluation that we present for the first time here—attempted to assess the system's ability to semantically interpret sentences from an open corpus using processing stages 1–5. As a reminder, this covers all modules before Situational Reasoning. Constraining the scope to non-situational semantics was necessary because neither our agents, nor any others within the current state of the art, have sufficiently broad and deep knowledge to engage in open-domain situational reasoning. The evaluations were carried out using a portion of the COCA corpus (Davies, 2008–). These experiments, like previous ones, required the agent to select those examples that it thought it could treat correctly.

The biggest challenges in applying our NLU engine to the open domain are incomplete coverage of the lexicon and incomplete coverage of our microtheories. These limitations are key to understanding the evaluation processes and outcomes, so let us consider them in more detail.

Incomplete lexicon. As a reminder, the lexicon that LEIAs currently use contains approximately 30,000 senses, which include individual words, multiword expressions, and constructions. This size is substantial for a deep-semantic, knowledge-based system, but it is still only a fraction of what is needed to cover English as a whole. In formulating our first holistic experiment, we attempted to account for this limitation by having the system select sentences that seemed to be fully covered by the lexicon. That is, we made the clearly oversimplifying assumption that if the lexicon contained the needed word in the needed part of speech, then there was a good chance that the needed sense was among the available options.

This assumption turned out to be false more often than anticipated, but it wasn't completely unfounded. The knowledge engineers who acquired the bulk of the lexicon some two decades ago were instructed to embrace, rather than back away from, ambiguity. And, in fact, the lexicon amply represents ambiguity. On average, prepositions have three senses each, conjunctions have three, and verbs have two. Ninety-eight verbs have more than five senses each, and the light verbs *make* and *take* have over forty and thirty senses, respectively. Nouns, adjectives, and adverbs average slightly over one sense each. The fact that not *all* senses of all words were acquired from the outset reflects competing demands on acquisition time, not an intentional avoidance of ambiguity. After all, for an open-domain lexicon (unlike a lexicon crafted for a narrowly defined application), there is no advantage to omitting word senses that have a reasonable chance of appearing in input texts.

The fact that a lexicon can contain a lot of senses but still lack the one(s) needed was amply demonstrated in these experiments. Consider just one example. The lexicon contains nineteen senses of *turn*, covering not only the core, physical senses (rotate around an axis and cause to rotate around an axis) but also a large number of multiword expressions in one or more of their senses: for example, *turn in*, *turn off*, *turn around*, *turn away*. Each of these is provided with syntactic and semantic constraints to enable automatic disambiguation. However, although nineteen well-specified senses sounds pretty good, our experiment used practically none of these and, instead, required three senses that the lexicon happened to lack:

- *Turn to*, meaning 'to face (physical)': *She turns to Tripp.* (COCA)
- *Turn to*, meaning 'to seek emotional support from': *People can turn to a woman.* (COCA)
- *Turn to food*, meaning 'overeat in an attempt to soothe one's emotions': *I'd always turn to food.* (COCA)

This means that lexicon lookup is not a reliable guide for determining whether a given input is or is not treatable. Relying on lexicon lookup is tantamount to a child's overhearing a conversation about parse trees and assuming that the trees in question are the big leafy things. The upshot is that the system often thinks it is getting the answer right when, in fact, it is mistaken. Later we will return to the important consequences of this both for system evaluation and for lifelong learning by LEIAs.

Incomplete coverage of microtheories. The second coverage-related complication of holistic evaluations involves microtheories. As readers well understand by now, although our microtheories attempt to cover all *kinds* of linguistic phenomena, they do not yet cover all *realizations* of each one—that will require more work.[10] In our first holistic experiment, we did not directly address the issue of incomplete coverage of microtheories. This resulted in the system's attempting to analyze—and then analyzing incorrectly—inputs containing realizations of linguistic phenomena that we knew were not yet covered.

So, for our second holistic evaluation, we improved the example selection process by formalizing what each microtheory did and did not cover, and we used this knowledge to

create a set of sentence extraction filters. This added a second stage to the task of selecting sentences to process as part of the evaluation. First the system extracted sentences that seemed to be covered by the lexicon. Then it filtered out those that contained phenomena that our microtheories do not yet cover.

These filters are not just an engineering hack; they are the beginning of a *microtheory of language complexity*. We do not call it *the* microtheory of language complexity because it reflects a combination of objective linguistic reality and idiosyncratic aspects of our environment.[11] The full inventory of extraction filters combines unenlightening minutiae that we will not report with points of more general interest, which we describe now.

The *intrasentential punctuation mark filter* rules in sentences with intrasentential punctuation marks that are either included in a multiword expression (e.g., *nothing ventured, nothing gained*) or occur in a rule-in position recorded in a list (e.g., commas between full clauses, commas before or after adverbs). It rules out sentences with other intrasentential punctuation marks, which can have a wide variety of functions and meanings, as illustrated by (9.17)—(9.19).

(9.17) She squeezed her eyelids shut, damming the tears. (COCA)

(9.18) Working light tackle, he had to give and take carefully not to lose it. (COCA)

(9.19) Now, think, she thought. (COCA)

The *relative spatial expression filter* excludes sentences containing relative spatial expressions because their meanings (e.g., *to the far left of the table*) can only be fully grounded in a situated agent environment. We are currently developing the associated microtheory within a situated agent environment, not as an exclusively linguistic enterprise.

The *set-based reasoning* and *comparative* filters exclude complex expressions that require constructions that are not yet covered in the lexicon.

(9.20) The second to last thing she said to him was, …. (COCA)

(9.21) In these stories he's always ten times smarter than the person in charge. (COCA)

The *conditional filter* rules in conditionals whose *if*-clause uses a present-tense verb and no modality marker (the *then*-clause can contain anything). It rules out counterfactuals, since counterfactual reasoning has not yet made it to the top of our agenda.

The *multiple negation filter* excludes sentences with multiple negation markers since they can involve long-distance dependencies and complex semantics.

(9.22) Except around a dinner table I had never before, at an occasion, seen Father not sit beside Mother. (COCA)

The *no-main-proposition filter* detects nonpropositional sentences that must necessarily be incorporated into the larger context.

(9.23) As fast as those little legs could carry him. (COCA)

(9.24) Better even than Nat and Jake expected. (COCA)

Note that the system *can* process the latter when it has access to multiple sentences of context, but in the reported experiments it did not.

The *light verb filter* excludes some inputs whose main verb is a light verb: *have*, *do*, *make*, *take*, and *get*. Specifically, it rules in inputs that are covered by a multiword expression that uses these verbs, and it rules out all others. The reason for this filter is that we know that the lexicon lacks many multiword expressions that contain light verbs. And, although the lexicon contains a fallback sense of each light verb that can formally treat most inputs, the analyses generated using those senses are often so much vaguer than the meaning intended by the input that we would evaluate them as incorrect.

For example, in our testing runs, use of the fallback sense led to overly vague interpretations of *take a cab*, *make the case*, and *get back to you*, all of which are not fully compositional and require their own multiword lexical senses. Note that this exclusion is not actually as strict as it may seem because constructions that the lexicon does contain actually cover large semantic nests. For example *have* + NP_{EVENT} means that the subject is the AGENT of the EVENT, which handles inputs like *have an argument*, *have an affair*, and *have a long nap.*

Let us pause to recap where we are in our story of holistic evaluations. Both holistic evaluation experiments encountered the same problem related to lexical lacunae: the lexicon could contain the needed word in the needed part of speech, but not the needed sense (which was often part of a multiword expression). As concerns the coverage of microtheories, the first experiment made clear that we needed to operationalize the agent's understanding of what each microtheory did and did not cover. We did that for the second experiment using the kinds of sentence extraction filters just illustrated.

In the first holistic experiment, which did not use the microtheory-oriented sentence extraction filters, there was a high proportion of difficult sentences that were beyond the system's capabilities. Some of the problems reflected how hard NLU can be, whereas others pointed to suboptimal decisions of experimental design. Starting with the problem that NLU is very difficult, consider the following sets of examples, whose challenges are described in brackets. As applicable, constituents of interest are italicized.

(9.25) [Compositional analysis failed due to a multiword expression not being in the lexicon.]

a. She is *long gone from* the club. (COCA)
b. I *got a good look at* that shot. (COCA)
c. The Knicks *can live with* that. (COCA)
d. But once Miller *gets on a roll*, he can make shots from almost 30 feet. (COCA)
e. I *can't say enough about* him. (COCA)
f. *This better be good*. (COCA)
g. You *miss the point*. (COCA)

(9.26) [A nonliteral meaning was intended but not detected.]
He not only hit the ball, he *hammered* it. (COCA)

(9.27) [It would be difficult, even for humans, to describe the intended meaning given just the single sentence of context.]
a. Training was a way of killing myself without dying. (COCA)
b. The supporting actor has become the leading man. (COCA)
c. This is about substance. (COCA)
d. The roots that are set here grow deep. (COCA)

(9.28) [The intended meaning relies more on the discourse interpretation than on the basic semantic analysis.]
a. It takes two to tango. (COCA)
b. And he came back from the dead. (COCA)

(9.29) [It is unclear what credit, if any, to give to a basic semantic interpretation when a large portion of the meaning involves implicit comparisons, implicatures, and the like.]
a. She's also a woman. (COCA)
b. How quickly the city claimed the young. (COCA)
c. They sat by bloodline. (COCA)
d. I think he is coming into good years. (COCA)
e. Fathers were for that. (COCA)

(9.30) [Without knowing or inferring the domain—the examples below refer to sports—it is impossible to fully interpret some utterances.]
a. The Rangers and the Athletics have yet to make it. (COCA)
b. He hit his shot to four feet at the 16th. (COCA)
c. We stole this one. (COCA)
d. I wanted the shot. (COCA)

As concerns the experimental setup for the first holistic evaluation, two of our decisions were suboptimal. First, we required the TMR for the entire sentence to be correct, which was too demanding. Often, some portion nicely demonstrated a particular functionality, while some relatively less important aspect (e.g., the analysis of a modifier) was wrong. Second, we focused exclusively on examples that returned exactly one highest-scoring TMR candidate (for practical reasons described below). We did not consider cases in which multiple equally plausible candidates were generated—even though this is often the correct solution when sentences are taken out of context. For example, the system correctly detected the ambiguity in, and generated multiple correct candidates, for (9.31) and (9.32).

(9.31) [The fish could be an animal (FISH) or a foodstuff (FISH-MEAT).]
He stared at the fish. (COCA)

(9.32) [*Walls* could refer to parts of a room (WALL) or parts of a person undergoing surgery (WALL-OF-ORGAN)]
He glanced at the walls. (COCA)

There are two reasons—both of them practical—for excluding sentences with multiple high-scoring candidate interpretations. First, since sentences can contain multiple ambiguous strings, the number of TMR candidates can quickly become large and thus require too much effort to manually review. Second, we would have needed a sophisticated methodology for assigning partial credit because, not infrequently, some but not all of the candidates are plausible.

Despite all the linguistic complications and tactical insufficiencies, our first holistic experiment yielded quite a number of satisfactory results, as shown by the following classes of examples.

(9.33) [Many difficult disambiguation decisions were handled properly. For example, this required disambiguating between sixteen senses of *look*.]
He looked for the creek. (COCA)

(9.34) [Many highly polysemous particles and prepositions were disambiguated correctly.]
a. She rebelled against him. (COCA)
b. He stared at the ceiling. (COCA)

(9.35) [Modification (*old*, *white*) and sets (*couple*) were treated properly.]
An old white couple lived in a trailer. (COCA)

(9.36) [Multiword expressions were treated properly.]
He took me by surprise. (COCA)

(9.37) [Dynamic sense bunching allowed the system to underspecify an interpretation rather than end up with competing analyses. For *ask* the system generalized over the candidates REQUEST-INFO, REQUEST-ACTION, and PROPOSE, positing their closest common ontological ancestor, ROGATIVE-ACT, as the analysis.]
I didn't ask him. (COCA)

(9.38) [New-word learning functioned as designed: the unknown word *uncle* was learned to mean some kind of HUMAN since it filled the AGENT slot of ASSERTIVE-ACT.]
The uncle said something to him. (COCA)

Turning to the second holistic experiment, it was different in three ways, the second of which (the use of filters) was already discussed.

1. *Example extraction*. The system sought examples of each verbal sense in the lexicon, still requiring that all other words used in the sentence be covered by the lexicon as well. This bunched results for easier comparative review.

2. *Filters*. We implemented the sentence extraction filters described earlier to automatically weed out linguistic phenomena that were known to not yet be covered by our microtheories.
3. *Defining "correct."* We developed a more explicit definition of a correct TMR, such that a correct TMR represented *a* (possibly not the only available) correct interpretation. In assessing correctness, we tried hard not to allow ourselves to question every decision knowledge engineers made when building the lexicon. For example, all words expressing breeds of dogs are mapped to the concept DOG since we never concentrated on application domains for which distinctions between dog breeds were important. So, a TMR that analyzed *poodle* as DOG would be considered correct. In short, our definition of *correct* allowed for underspecifications deriving from knowledge-acquisition decisions, but it did not allow for actual mistakes. If the lexicon was lacking a needed word sense, the fact that the agent used the only sense available does not make it right. After all, this is an evaluation of semantic analysis; it is not code debugging.

As in the first holistic evaluation, we excluded sentences that offered multiple high-scoring analyses, and we evaluated sentence-level analyses on the whole. The rationale was as before: to avoid introducing excessive complexity into the evaluation setup.

Before turning to the successes of the second holistic evaluation, let us consider some failures. No surprise: most of them were due to missing lexical senses, including multiword expressions and constructions. Here are just a few examples for illustration:

(9.39) [The lexicon contains the direct/physical meaning of the underlined word but not the conventional metaphor, which also must be recorded.]
a. Christians in Egypt worry about the ascent of Islamists. (COCA)
b. Shaw had won the first battle. (COCA)

(9.40) [Although the lexicon contains several nominal senses of *line*, *clothesline* was not among them.]
They wash their clothes, and they hang them on a line. (COCA)

(9.41) [The system found a sense of *wait out* in the lexicon, but that sense expected the complement to be an EVENT (e.g., *wait out the storm*), not an OBJECT. When the complement is an OBJECT, the expression is actually elliptical: the named OBJECT is the AGENT of an unspecified EVENT whose meaning must be inferred from the context. Our lexicon already contains treatments of lexemes requiring a similar ellipsis detection and resolution strategy (e.g., covering *A raccoon caused the accident*); it just happens to lack the one needed for this input.]
They waited out the bear. (COCA)

(9.42) [*Would always* is a multiword expression indicating that the event occurred repeatedly in the past. The lexicon lacked this multiword sense at the time of the evaluation run.]
My mother would always worry. (COCA)

(9.43) [This entire sentence is idiomatic, conveying a person's inability to think of the precise word(s) needed to express some thought. Analyzing it compositionally is just wrong.]
I wish I had the word. (COCA)

In many cases, the meaning of a sentence centrally required understanding implicatures. That is, the sentence did not merely give rise to implicatures; instead, arriving at the basic meaning required going beyond what was stated. We did not give the system credit for implicature-free interpretations in such cases, even if they contained correct aspects of the full meaning.

(9.44) [This does not simply mean that individuals representing the IRS will serve as collaborating agents with *them* on something. It means working out a way for the people to pay off their tax burden to the IRS.]
The IRS will work with them. (COCA)

(9.45) [This sentence does not involve a single instance of writing a single sentence, as the basic analysis would imply. Instead, it means that he is an excellent writer, an interpretation that relies on a construction (cf. *She plays a mean horn*; *He makes a delicious pizza pie*).]
He writes a great sentence. (COCA)

(9.46) [Neither of the auxiliary senses of *can* in the lexicon (indicating ability and permission) is correct for this sentence. Here, *can* means that they have written, and have the potential to write in the future, such emails.]
They can also write some pretty tough e-mails. (COCA)

(9.47) [This elliptical utterance requires the knowledge that *Parks and Recreation* is a department that is part of the city government.]
I work for the city, Parks and Recreation. (COCA)

(9.48) [This implies that golfers do not ride in golf carts, not that they simply walk around in principle.]
Golfers have always walked in competitive tournaments. (COCA)

(9.49) [A full interpretation requires identifying which features of a felon are salient.]
I was no better than a felon. (COCA)

(9.50) [This refers to particular political actions (perhaps protesting or contacting voters), not strolling around.]
I will walk for candidates. (COCA)

(9.51) [Compositionally, this means that members of the cabinet are coagents of voting. However, there is a political sense—relevant only if the indicated people represent appropriate political roles—that means *to vote the same as a higher-positioned politician*.]
The cabinet voted with Powell. (COCA)

It should be clear by now why counting things is a poor yardstick for evaluation. There's something not quite fair about marking an open-domain system wrong for not inserting golf carts into the interpretation of *Golfers have always walked in competitive tournaments*—especially when many human native speakers of English probably don't know enough about golf to understand what was meant either.

So, as in the first holistic evaluation, in the second one we oriented around qualitative rather than quantitative analysis, focusing on (a) what the system *did* get right, as proof of concept that our approach and microtheories are on the right track, and (b) lessons learned.

We halted the experiment after collecting fifty sentences that the system processed correctly, since by that point we had learned the big lessons and found ourselves just accumulating more examples of the same. Those fifty sentences are listed below, with selective, highly abbreviated comments on what makes them interesting.

(9.52) [*Told* was disambiguated from six senses of *tell*.]
Shehan told him about the layoffs. (COCA)

(9.53) [There are multiple propositions.]
Dawami says neighbors told her they heard Hassan beat the girl. (COCA)

(9.54) [There are multiple propositions and volitive modality from *hope*.]
I told him I hope he wins. (COCA)

(9.55) [The TMR is explanatory: TEACH (THEME INFORMATION (ABOUT BUDDHISM))]
Monks teach you about Buddhism. (COCA)

(9.56) [*Poetry* is described as LITERARY-COMPOSITION (HAS-STYLE POETRY). *Write* was disambiguated from eight senses.]
I write poetry. (COCA)

(9.57) [*Worsen* is described as a CHANGE-EVENT targeting the relative values of evaluative modality in its PRECONDITION and EFFECT slots.]
You'd worsen the recession. (COCA)

(9.58) [*About* was disambiguated due to the inclusion of the multiword expression *worry about*.]
I worried about him. (COCA)

(9.59) [Turin was correctly analyzed as CITY (HAS-NAME 'Turin').]
He worried about Turin's future. (COCA)

(9.60) [*On* was disambiguated due to the inclusion of the multiword expression *work on*.]
I'll work on the equipment. (COCA)

(9.61) [*And* creates a DISCOURSE-RELATION between the meanings of the propositions. The instances of *we* are coreferred.]
We work and we eat. (COCA)

(9.62) [*Fast* is analyzed as (RAPIDITY .8).]
I work fast. (COCA)

(9.63) [*Championship* is an unknown word that was learned as meaning some sort of EVENT.]
The team won the championship. (COCA)

(9.64) [The multiword sense for *watch out* is used. There is obligative modality from *should*.]
Everybody should watch out. (COCA)

(9.65) [The causative sense of *wake up* is correctly integrated with the obligative modality from *should*.]
They should wake you up. (COCA)

(9.66) [The instances of *I* are correctly coreferred. *Wake up* is described using the *end* value of ASPECT scoping over a SLEEP event whose EXPERIENCER is the HUMAN indicated by *I*.]
I slept till I woke up. (COCA)

(9.67) [*Wake up* is analyzed as above. *With* is correctly interpreted as BESIDE.]
Jack wakes up with Jennifer. (COCA)

(9.68) [The conjunction structure is analyzed as a set. The proper names are correctly analyzed as two cities and a state with their respective names.]
He has visited Cincinnati, Tennessee, and Miami. (COCA)

(9.69) [The analysis explicitly points to all the concepts relevant for reasoning: COME (DESTINATION (PLACE (LOCATION-OF CRIMINAL-ACTIVITY))).]
They visited the crime scenes. (COCA)

(9.70) [The modification and nominal compound are treated correctly.]
Ghosts visit a grumpy TV executive. (COCA)

(9.71) [The multiword expression *turn down* allows the system to disambiguate among twenty-six verbal senses of *turn*.]
Thoreen turned down the offer. (COCA)

(9.72) [This sense of *use* is underspecified, instantiating an EVENT whose INSTRUMENT is the set ONION, GARLIC, CUMIN. Although people would probably infer that seasoning food is in question, this is not a necessary implicature: this sentence could also refer to gardening or even painting.]
You use onion, garlic, and cumin. (COCA)

(9.73) [This uses the multiword expression *turn off.*]
He turns off the engine. (COCA)

(9.74) [This uses the multiword expression *turn on*.]
Somebody turned on a television. (COCA)

(9.75) [The question is interpreted as a request for information: namely, the agent of the proposition.]
Who trained them? (COCA)

(9.76) [Two different senses of *agent* work here: an intelligence agent and the agent of an event more generically. This instance of residual ambiguity launched an automatic sense-bunching function that generalized to their most common ancestor, HUMAN. It left a trace of that generalization in the TMR, in case the agent later chooses to seek a more specific interpretation using discourse-related reasoning.]
We train our agents. (COCA)

(9.77) [This uses the multiword expression *track down*. Also note that the lexicon acquirer chose to attribute null semantics to *always* because it rarely, actually, means *always*! For example, *He is always teasing me* does not literally mean all the time. One can disagree with this acquisition decision, but it was a conscious, documented decision that we did not overturn for this experiment.]
Chigurh always tracks him down. (COCA)

(9.78) [This uses *belief* modality from *think* and the multiword expression *find out*.]
I think we found out. (COCA)

(9.79) [This uses the multiword expression *think about*.]
They think about the road. (COCA)

(9.80) [*For* was correctly disambiguated (from among eighteen senses) as PURPOSE.]
We stayed for lunch. (COCA)

(9.81) [*Forever* is analyzed as 'TIME-END never'.]
Joseph would stay there forever. (COCA)

(9.82) [This uses the MWE *stand up for* and the property ASPECT with the value *end* scoping over the main event, PROTECT (from *stand up for*).]
Maeda had stood up for Mosley. (COCA)

(9.83) [*Speed up* is described as a CHANGE-EVENT whose PRECONDITION and EFFECT have different relative values of SPEED.]
They speed up. (COCA)

(9.84) [This uses the multiword expression *sign off on*.]
The courts must sign off on any final accounting. (COCA)

(9.85) [*Showered* is described in the lexicon as BATHE-HUMAN (INSTRUMENT SHOWER).]
He showered. (COCA)

(9.86) [This uses the multiword expression *shoot back at*.]
Nobody had shot back at them. (COCA)

(9.87) [The nominal compound *lab space* is analyzed using a generic RELATION since the candidate meanings of the nouns do not match any of the more narrowly defined ontological patterns supporting compound analysis.]
They share lab space. (COCA)

(9.88) [*Policy* is described in the lexicon as a necessary procedure—that is, PROCEDURE scoped over by obligative modality with a value of .7]
They would shape policy. (COCA)

(9.89) [*Settle in* is correctly disambiguated as INHABIT.]
They settled in Minsk. (COCA)

(9.90) [*Never* is described using epistemic modality with a value of 0 scoping over the proposition.]
He'll never send the money. (COCA)

(9.91) [*See* is disambiguated from thirteen available senses.]
I'll see you on the freeway. (COCA)

(9.92) [Highly polysemous *see* and *at* are correctly disambiguated; there is modification of a proper noun and interpretation of a nominal compound (*cafeteria door*).]
She saw an injured Graves at the cafeteria door. (COCA)

(9.93) [*Legislator* is described in the lexicon as POLITICIAN (MEMBER-OF LEGISLATIVE -ENTITY).]
Legislators scheduled hearings. (COCA)

(9.94) [*Ran* is disambiguated from twelve available senses.]
I ran to the door. (COCA)

(9.95) [This uses the MWE *run a campaign*; *but* instantiates the discourse relation CONTRAST between the meanings of the clauses; the lexical description of *prefer* uses evaluative modality; and the proper names are correctly handled.]
Biss has run a good campaign, but we prefer Coulson. (COCA)

(9.96) [This uses the multiword expression *rise up*.]
We will rise up. (COCA)

(9.97) [This uses the lexical sense for the middle voice of *ring*.]
A dinner bell rang. (COCA)

(9.98) [This uses the multiword expression *station wagon*.]
We rented a station wagon. (COCA)

(9.99) [*Release* is correctly analyzed as INFORM.]
The NCAA releases the information. (COCA)

(9.100) [*Refuse* is analyzed as an ACCEPT event scoped over by epistemic modality with a value of 0—that is, *refusing* is described as *not accepting*.]
USAID refused interviews with staff in Badakhshan. (COCA)

(9.101) [*Race* is correctly understood as a MOTION-EVENT with a VELOCITY of .8 (not an actual running race).]
She raced to the church. (COCA)

The automatically generated TMRs for these examples are available at https://homepages.hass.rpi.edu/mcsham2/Linguistics-for-the-Age-of-AI.html.

This pair of holistic experiments served its purposes. First, they validated that the system was working as designed and could generate impressive analyses of real, automatically selected inputs from the open domain. Second, they highlighted the need for the *microtheory of language complexity* and led to our developing the first version of that microtheory. Third, they gave us empirical evidence that has allowed us to make a major improvement in our system: a redesign of confidence assessment for TMRs.

Before these experiments, our confidence measures relied exclusively on how well the input aligned with the syntactic and semantic expectations recorded in our knowledge bases. However, as we have seen, when an analysis seems to work fine, the agent can fail to recognize that it is missing a word sense, multiword expression, construction, or piece of world knowledge needed for making implicatures. So we now understand that it is important to enable the system to do all of the following:

1. When operating in a particular application area (i.e., a narrow domain), the LEIA will need to distinguish between in-domain and out-of-domain utterances. As a first approximation, this will rely on frequency counts of words describing concepts participating in known ontological scripts.
2. The LEIA will need to apply additional reasoning to in-domain utterances, effectively asking the question, "Could the input have a deeper or different meaning?"
3. The LEIA will need to decrease overall confidence in out-of-domain analyses due to the fact that they are not being fully semantically and pragmatically vetted using the kinds of ontological knowledge a person would bring to bear.

All of these have important implications for lifelong learning, which is a core functionality if agents are to both scale up and operate at near-human levels in the future. That is, although an agent can learn outside its area of expertise, that learning will be of a different quality than learning within its area of expertise. This suggests that the most efficient approach to learning will involve starting from better-understood domains and expanding from there. It is worth noting that, although we have been in this business for a long time, we would not have realized how frequently this overestimation of confidence occurs—that is, how frequently the agent thinks it has understood perfectly when, in fact, it has not—had we not gone ahead and developed a system and evaluated it over unrestricted text. There are just no introspective shortcuts.

9.4 Final Thoughts

One of the strategic decisions used in all the reported experiments (the five devoted to individual microtheories and the two holistic ones) was to challenge the system with examples

from the open domain but allow it to select the examples it believed it could treat effectively. The rationale for this independent-selection policy is that we are developing intelligent agents that will need to be able to collaborate with people whose speech is not constrained. This means that utterances will be variously interpretable. For example, a furniture-building robot will lose the thread if its human collaborators launch into a discussion of yesterday's sports results. So, given each input, each agent must determine what it understands and with what confidence. This is the same capability that each evaluated subsystem displayed when it selected treatable examples from an open corpus. It reflects the agent's introspection about its own language understanding capabilities.

As we saw, the biggest hurdle in correctly making such assessments—and our biggest lesson learned—involves the lexicon. It can be impossible for an agent to realize that it is missing a needed lexical sense (which might be a multiword expression or construction) when the analysis that uses the available senses seems to work fine. Three directions of R&D will contribute to solving this problem.

1. Redoubling our emphasis on learning by reading and by interaction with humans, which is the most practical long-term solution for resource acquisition. The best methodology will be to start with domains for which the agent has the most knowledge—and, therefore, can generate the highest-quality analyses—and spiral outward from there.
2. Consulting lists of potential multiword expressions (which can be generated in-house or borrowed from statistical NLP) during language processing. These lists will contain (potential) MWEs that are not yet recorded in our lexicon and, therefore, are not yet provided with semantic interpretations. However, they will serve as a red flag during processing, suggesting that the compositional analysis of the given input might not be correct. This should improve our confidence scoring, helping the agent to not be overconfident in analyses that might not be fully compositional.
3. Carrying out manual lexical acquisition. Although manual acquisition is too expensive to be the sole solution to lexical lacunae, it would be rash to exclude it from the development toolbox, particularly since it is no more time-consuming than many other tasks that are garnering resources in the larger NLP community, such as corpus annotation.

As should be clear, our agents can work in various modes. In application mode, they must do the best they can to process whatever inputs they encounter. In component-evaluation mode, they attempt to identify inputs that they believe they can interpret correctly. And in learning mode, they use their ability to assess what they do and do not understand to identify learnable information.

To return to the starting point of this chapter, there is no simple, all-purpose strategy for evaluating knowledge-based systems. Crafting useful, feasible evaluation suites is an ongoing research issue. Evaluations are useful to the extent that they teach us something that we would not have understood through introspective research practices and normal test-and-debug cycles. It is, therefore, not a stretch to say that bad evaluation results can be a blessing in disguise—as long as they lead to new insights and suggest priorities for future R&D.