Broad-domain Quantifier Scoping with RoBERTa

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

in the Graduate School of The Ohio State University

By

Nathan Ellis Rasmussen, M.A.

Graduate Program in Linguistics

The Ohio State University

2022

Dissertation Committee:

Professor William Schuler, Advisor

Professor Micha Elsner

Professor Michael White

Abstract

This thesis reports development of a new, broad-domain quantifier scope corpus including all of the factors, for use training and testing the system. Training materials, a work process, and the annotator-facing data format were each designed to reduce barriers to entry and safeguard accuracy, with revisions resulting from an inter-annotator agreement study and error analysis.

The thesis discusses appropriate measures of agreement for scope annotations, both between human annotators and between predicted and gold labels. For appropriate calculation of chance-corrected agreement between human annotators, an inter-annotation distance metric is introduced and justified. For evaluation of automated predictions, where human-like constraints on the structure of a set of predictions are not enforced, results are evaluated both for small-scale accuracy and for compliance with these holistic constraints.

The scoping data of the corpus are developed into a natural language understanding task suitable for automatic prediction, framing it as a span pair classification problem, with outscoping treated as a semantic dependency between words.

This thesis reports the application of the RoBERTa language model to this task. The model encodes properties of lexis, syntax, and semantics that correlate with human scoping judgements ('scoping factors'). Previously published scope-annotated corpora and scope prediction systems either do not cover all of the scoping factors, do not apply them to the full set of quantifiers, or do not represent the full range of subject-matter domains in which humans routinely predict quantifier scope.

Predictions from the RoBERTa system are shown to be more accurate than the majority-prediction baseline, to a degree not due to chance. The system successfully complies with the holistic constraints. The system's principal shortcomings are its relatively small improvement over the baseline, its dependence on some other system to screen pairs of scope-bearers for the presence of

scopal interaction, and the inability thus far of its architecture to serve as that screener. Further steps to address these are proposed.

Dedication

To your memory, sweet Mona.

Acknowledgements

My studies have been financially supported by a fellowship, by teaching work, by research funds, by a Covid-emergency fee waiver, and by family. As concerns the fellowship, I owe thanks to Carl Pollard, William Schuler, and Shari Speer for their advocacy. For the teaching, an unexpected delight, I owe the department for funds and logistics, Hope Dawson for good counsel, and everyone who built our weird, wonderful codebreaking curriculum before or beside me, notably Micha Elsner and Daniel Puthawala. Research funds came via William from National Science Foundation grants #1551313 and #1816891, but views expressed are my own, not necessarily the NSF's or even William's. The emergency fee waiver is courtesy of the graduate school. And the other revenues have come from the sundry employments of my wife, Shelley Denison, and the valued emergency backing of our parents: Sue and David Rasmussen, Joy and Brent Denison. Many thanks to all of you.

Thanks to William, Micha, and Michael White for shepherding this work as my committee. William in particular, for being very much the kind of advisor I needed. Thanks also to annotators Tahiira Zimmerman, Rachel Meyer, and Juliette Rike for building up the data in spite of being regularly plunged headfirst through all manner of semantic arcana, and to everyone who shared their computing power in the Unity cluster.

Well-being, of course, is not just a matter of the pocketbook. This department's collegiality has meant the world to me, and I owe thanks to everyone who has helped to make it welcoming and humane—certainly in aspects like allowing for struggles and mishaps, but in truth no less from simple, casual friendliness that opened the way for many a valued conversation. There is no really fair place to draw a line between people whose company brought 'enough' goodness into my life to thank by name and those who 'did not'; so although I will name some that have come to mind,

really if you ever taught me anything, asked me to teach you anything, confided in me, or lent me an ear in the (rather long) time I've been at OSU, your name should be here.

Thanks to Julia Papke for games played, institutional expertise deployed, and for everything to do with our common background. In the area of games let me also mention Micha and especially Jim Harmon, who with Julia introduced me to some excellent ones. Thanks to Evan Jaffe, by my side more often than anyone as we worked our way through; to Rachel and James Burdin, Shelley's and my curious mirror image; to Marten van Schijndel, Manjuan Duan, Shuan Karim, Yourdanis Sedarous, Jordan Needle, Julie McGory, and Lifeng Jin; to the members of the CoCoMo lab and other frequent CaCLers; and to Brian Joseph, Don Winford, Craige Roberts, and Bridget Smith, whose warmth toward me has been vastly more than I'd ever have predicted from the extent of my involvement with them.

Thanks to the folk and especially the leaders of side interests that refreshed me to press on with this main line: Hope and Daniel in teaching, Hope again and the History of Sanskrit crew, Dan Collins and the Prehistory of Slavic bunch, Laura Wagner and the pod people. Additional citations of merit for Dan Collins helping to knock mental health stigma down and Laura Wagner for caring followup months after a rather gnarly physical health problem kicked my whole life apart.

Finally, thanks to all of my 'home people', of whatever species, for all the joys of knowing them and being theirs.

Vita

2005 ..... B.A. Linguistics
magna cum laude and with University Honors
Brigham Young University

2009 ..... M.A. Linguistics
Brigham Young University

2012–13 and 2020–21 ..... Distinguished University Fellow
The Ohio State University

2013–20 ..... Graduate Teaching and Research Associate
Department of Linguistics
The Ohio State University

Publications

Nathan Ellis Rasmussen and William Schuler. 2020. A corpus of encyclopedia articles with logical forms. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1051–1060.*

Micha Elsner, Andrea D Sims, Alexander Erdmann, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, David L King, Luana Lamberti Nunes, Byung-doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V Dickinson, and Michelle Mckenzie. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modeling,* 7(1):125—170.

Nathan E Rasmussen and William Schuler. 2018. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science,* 42(54):1009–1042.

Nathan Ellis Rasmussen and Deryle Lonsdale. 2012. Lexical correspondences between the Masoretic Text and the Septuagint. In *Proceedings of LREC'2012 Workshop LRE-Rel: Language Resources and Evaluation for Religious Texts,* pages 58–64.

Fields of Study

Major Field: Linguistics

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

Explanatory generalizations in natural language represent extensive experience in a compact and wieldy form, greatly accelerating human learning and enabling complex cultural development. Generalizations characteristically include quantificational noun phrases, often several of them in a simple sentence like Example (1).

(1) Most people have two hands with five fingers.

Where multiple quantifiers occur, there are often multiple distinct readings of the sentence or discourse, some of which may be true and others false at the same time, depending on the facts of the world. Example (1) could be taken to mean that or even that more than half of the population are in joint possession of the same five fingers, among other possibilities. Which meaning we reach depends on whether we evaluate the claim about *five fingers* inside or outside the claims about *two hands* and *most people,* a consideration known as scope.

A human reader will generally have little trouble recognizing the intended meaning, with different hands for each person and different fingers for each hand, if the ambiguity is even noticed at all. But as a computational task, predicting the preferred reading has been quite challenging, and systems that succeeded in it have been narrowly limited as to the quantifiers, syntactic structures, or subject-matter domains they handled.

Recently, powerful new tools have emerged for natural language processing, with the BERT system of Devlin et al. (2018) often cited as a turning point. BERT greatly expanded the utility of contextualized word embeddings, which are computed representations of a word's meaning and use

in its context, and this allowed eleven diverse natural language tasks[1] to share a majority of their architecture and training time, while each reaching a new state of the art.

Contextualized word embeddings incorporate multiple kinds of information known to correlate with human scoping judgements: the lexical realization of quantifiers (e.g. whether a universal is expressed as *every* or *each*), the linear order in which the quantifiers appear, the syntactic structure around them, the sense in which ambiguous words are used, the nature of the objects and events under discussion, and the anaphoric, inferential, and rhetorical connections between sentences in a discourse (including coherence and coreference) (Alshawi, 1992; AnderBois et al., 2012; Dwivedi, 2013; Kuno, 1991).[2] No previous quantifier scope disambiguation system has made use of all of these scoping factors. This, and the BERT approach's remarkable success in and after its debut, suggested these word embeddings might bring new power to the scope prediction task. Below, Section 1.2 gives further background on the nature and use of contextualized word embeddings, and Section 1.3 reviews the indications for and against their potential to predict scope.

1.1 Summary of thesis

This thesis applies a BERT-derived word embedding system to a set of scoping problems that exceed previous tasks' diversity of quantifiers, of syntactic environments in which they appear, and/or of subject matter. Unlike previous tasks, they are set in naturally produced connected texts, and so the quantifiers are are found in a wider variety of information structures than out-of-the-blue sentences usually offer, and their interpretation can be shaped by discourse relations among sentences.[3] The

[1] Including grammaticality and sentiment judgements, semantic similarity scoring and paraphrase detection, question answering, and inferential reasoning

[2] AnderBois et al. summarize from Micham et al. (1980); Fodor (1982); Gillen (1991); Kurtzman and MacDonald (1993); Tunstall (1998); Anderson (2004).

[3] Moreover, the scoping problems are set in texts whose purpose is only met if they accurately yet briefly communicate generalizations about complex subjects. The tension between accuracy and brevity demands a careful balance between building up knowledge through the text and presuming on the reader to supply background. The balance is not always struck successfully—as of our 2014 dump of the wiki, the article about fire called it 'one of the most familiar examples of the chemical process of oxidation', then proceeded to explain that 'A person should never touch fire'—but the practical communicative purpose and the tension within it help to ensure that presumed and previously stated knowledge each come into play often.

predicted labels are more accurate than a baseline, and separate predictions are logically consistent with one another in spite of their potential to create paradoxical scoping cycles. Chapter 4 describes how the scope judgements are extracted from a corpus and framed as a natural language processing task; Chapter 5 discusses the correct way to evaluate the results, then does so.

Creating scoping problems to fully exercise word embeddings' potential requires appropriate source data: scope-annotated documents that contain many quantifiers, talk about many aspects of the world, and exhibit a natural range of variation on the scoping factors. No extant corpus of quantifier scope judgements met this requirement, so we[4] have built our own out of excerpts from the Simple English Wikipedia.[5] This entailed annotating documents myself, developing and revising guidelines, measuring inter-annotator agreement and analyzing disagreement, and training and advising additional annotation workers. The corpus and the annotation process are described in Chapter 2.

The inter-annotator agreement study is discussed separately. Scoping annotations in connected texts violate the assumptions of statistics used for inter-annotator agreement in previous scope corpora. Krippendorff's α is a more appropriate measure, but requires a distance metric suited to the annotations' nature and able to compare any two of them (even annotations of two different documents). Chapter 3 explains the problem and the α statistic, describes and justifies the distance metric I devised, and reports the findings. Although the distance metric is specialized for this annotation scheme, aspects of its design are relevant to other kinds of internally-structured data. The findings include an error analysis, which led to improvements in the annotation process.

Chapter 6 summarizes the work and briefly discusses ways forward.

1.1.1 A cognitive caveat

A note of caution must be sounded. Although the scoping factors were identified in psycholinguistic literature, this is not a psycholinguistic model.

[4]Herein, by default, 'we' means Dr. Schuler, myself, and when context so indicates, other annotators.
[5]https://simple.wikipedia.org/

Human knowledge of the world is situated, or rooted in experience, even though it is greatly supplemented by language-borne generalizations. Word embeddings begin with world knowledge derived entirely from counting collocations in documents. Then they are enriched with linguistic knowledge acquired through an even less human-like process, by an encoder not at all resembling a silicon Wernicke's area, as it reads every word of the document simultaneously. In short, RoBERTa is a Martian.

A Martian with a good grasp of how we imply and infer quantifier scope can still do useful work, of course, whether or not its internal processing is analogous to Earthlings'. And by showing us our linguistic behavior 'from the side', so to speak, it may even reveal things about it that we were unaware of.

This research is offered to establish landmarks and break trail toward natural language computing that can handle our full palette of scopal cues, and can cooperate with our quick intuitive judgements about this subtle, important facet of meaning. Using the psycholinguistically identified scoping factors will, I hope, help to build a more complete model of what we do, but it should not be mistaken for a model of how we do it.

1.1.2 On notation

Our semantic model of quantification is the Generalized Quantifier (Barwise and Cooper, 1981), but this document sometimes paraphrases them as first-order quantifiers for brevity.

1.2 Contextualized word embeddings and BERT

This section describes contextualized word embeddings generally, and the BERT-like varieties more particularly. The tale begins with uncontextualized word embeddings, perhaps better called vocabulary embeddings.

1.2.1 Vocabulary embeddings

Vocabulary embeddings such as GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) represent each word type as a point in a many-dimensional space, such that syntactic and semantic similarities between words are reflected in their spatial relationships. A widely known example from Word2Vec, treating each embedded point as the vector to that point from the origin/zero point of the space, is that the vector for 'king', minus the vector for 'man', plus the vector for 'woman', approximately equals the vector for 'queen'. Thus, the semantic structure of the vocabulary is embedded in the space.

This is possible because word meaning is reflected in word usage; 'you shall know a word by the company it keeps' (Firth, 1957). The settings in which 'king' appears are much like those for 'queen', except for their correlation with gendered language such as pronouns. Their statistical patterns of co-occurrence with other words therefore reflect the semantic relationship. Morphosyntactic facts like part of speech are also captured by the same statistics; words that we would call countable common nouns will tend to appear directly after 'the', 'a' or 'an', or various adjectives, and to appear directly before auxiliary verbs, main verbs, or postmodifiers such as relative pronouns and prepositions.

When the vast, diffuse statistics describing co-occurrence among several thousand word types are condensed to a few hundred more abstract values that adequately predict their distribution, the semantic and morphosyntactic patterning is retained, and the result is a vocabulary embedding. Each of the few hundred values locates a word along one dimension; knowledge about the semantic and morphosyntactic patterns is implicit in and distributed across the word's position on these dimensions and its spatial relationships with other words.

The technique can be extended in various reasonable ways. For example, subword-enriched vectors (Bojanowski et al., 2017) also provide embeddings for character *n*-grams within words, so that appropriate points can be assigned to words that were not seen in training but that are related to trained words (either by morphology or by misspelling).

In natural language processing tasks, using these vectors to represent words provides an injection of (usually helpful) information about their semantics and typical syntactic behavior. However, any embedding that provides a fixed representation of each word type suffers from homonymy and polysemy, which limit the amount of world knowledge a word can provide; and it cannot represent grammatical relationships between words in use, which are believed to affect human scoping decisions. Thus, as an information source for scope prediction (and for many other tasks), vocabulary embeddings are incomplete.

1.2.2 Contextualization and attention

Contextualized word vectors (for example, Melamud et al., 2016; McCann et al., 2017; Peters et al., 2017, 2018a; Devlin et al., 2018; Radford et al., 2018), represent word types' identities and meanings as vocabulary embeddings do, but refine the representation of each word token to embed information about its use in its own unique sentence context—its position relative to other words, the subset of its typical syntactic behaviors that are actually in use, the sense in which it is used, and so forth.

Generally this is done by training a neural network on a task that requires that information. It will begin with a text's vocabulary embeddings, impose a representation of their position in the text's sequence,[6] process them in a way that allows their values to be affected by the values of other words in the text, and use them for the task, so that the training process will teach the network to extract the additional semantic and syntactic details for each word from the words around it. After training, arbitrary texts can be read by the network as though to perform the task, and its internal representation of each word can be extracted as a contextualized embedding.

Various tasks can be used to induce the network to represent the additional information about use-in-context. Kawakami and Dyer (2015) take the approach that the additional information about word sense and syntax is implicit in a correct translation into another language, and so their word-in-context representation is the hidden state reached at a word by a neural machine translation system,

[6] Either explicitly by altering the vector, or implicitly by processing them in order with the help of a persistent memory.

which encodes one language into embeddings that it then decodes into another language. McCann et al. (2017) do much the same, but the translation system they train has an attention component. Attention (Bahdanau et al., 2014) uses the state of the output decoder to select words of the input that are particularly relevant for predicting the next output word, so that in effect the system is learning to predict dependencies between input and output words.

Melamud et al. (2016), on the other hand, conceive of contexts as the additional information needed for a good model of monolingual text. This approach has the advantage that monolingual text for training is much easier to obtain than quality translation pairs are, particularly if the training task is one like cloze completion that requires no additional annotator input. Language modeling has become the usual form of pretraining as subsequent work has explored various architectures for the trainee (Peters et al., 2017, 2018a; Devlin et al., 2018; Radford et al., 2018; Zhou and Srikumar, 2019).[7] Many of these models, including BERT, also use self-attention: attention for predicting dependencies among the words of the monolingual text, rather than between translation equivalents.

1.2.3 Fine-tuning

The quantity of training that goes into such models continues to grow. As they learn to make each word's embedding a useful predictor of other words for language modeling, the information they incorporate has indeed proven to be transferable to a wide variety of other language processing tasks. A particularly strong technique uses the (typically much smaller) task-specific data not just to train a task-specific neural network that uses the embeddings from the language modeling encoder, but to fine-tune the encoder itself.

It is generally thought (for example Devlin et al., 2018; Peters et al., 2018b; Tenney et al., 2019) that the encoder is able to discover more about inter-word dependencies than is strictly necessary for, say, cloze completion. However, the last few layers of its architecture, because they are structurally

[7] There is evidence (Wang et al., 2019) strongly suggesting that different tasks benefit to different degrees from language modeling or other forms of pretraining, and that language modeling alone may not suffice to produce truly general-purpose embeddings. However, it seems to be the most productive single task at the moment, and its data availability is unmatched.

closest to the task output, become specialized for the pretraining task. They therefore learn to filter out information that is, for that task's purposes, unduly noisy or complex.

Fine-tuning reworks these layers to better suit the new task, fishing up information that they formerly discarded and/or suppressing information the new task does not need. The early and middle layers of the encoder already have vast experience extracting the patterns of language; fine-tuning takes advantage of it.

BERT's original elevenfold success on a range of tasks was accomplished exactly this way. Ongoing successes, both with BERT and with other (usually attention-enabled) massively pretrained language modeling encoders, have led to the conclusion that "language models are unsupervised multitask learners" (Radford et al., 2019). Though that claim may be under-nuanced (see discussion in Appendix A), this is in any case good news for a task like scope prediction, where task-specific training data is scarce and older techniques have been taxed to their limits.

1.3 Feasibility

This section reviews the performance of BERT (and, to a lesser extent, other contextualized word embeddings, particularly others that are encoded with self-attention) on tasks judged to be similar in nature or difficulty to scope prediction, as of the date I evaluated and chose this approach.

Quantifier scope is generally described in terms of recursive structural embedding, either the recursion of predicate logic terms in the approach of Montague (1973), or that of covert syntactic constituents in the approach of May (1977). But if the structural pattern is abstracted away from some of the details of the material it structures, outscoping can take the appearance of a dependency, as it does in Schuler and Wheeler (2014). In fact, we can view it as a dependency between words as well as a structural relationship of the semantic objects they signify, just as syntax can be viewed in terms of dependencies between words as well as of the structural relationships between constituents they head.

From this perspective, BERT's potential for the outscoping task can be roughly gauged by its ability to encode other long-distance semantic relationships/dependencies, as it may do in tasks of coreference identification, summarization, or even clozing, if it is highly context-sensitive.

1.3.1 Overviews

This section contains discussion about the feasibility of the approach in general.

1.3.1.1 Optimistic overviews

In terms of the scoping factors where vocabulary embeddings fall short, contextualization appears to make up the difference: They support word sense disambiguation (Peters et al., 2018a), which improves access to world knowledge. They capture very substantial information about the syntactic environment of the words they represent (Kitaev and Klein, 2018; Peters et al., 2018b). And they carry discourse information suitable for resolving anaphora (Lee et al., 2018; Peters et al., 2018b). Identifying an outscoper is likely more difficult on average than identifying a coreferential antecedent, but from the perspective of scope-as-dependency it may be a difference of degree and not one of kind entirely. Thus, they may already be the right sort of representation for a quantifier prediction task.

1.3.1.1.1 Contextualization represents syntax thoroughly Several studies demonstrate that syntax is represented especially well in the contextualization. Mohammadshahi and Henderson (2019) used an attention-based system to generate a sequence of parsing actions and resulting syntactic dependencies. In addition to encoding the parser state, history of actions, and the dependents of the top words on the parse stack, they devised a way to encode the dependencies already created as an input to the attention function. The word encoder could be either trained from scratch, or initialized with pre-trained BERT parameters (no fine-tuning). With BERT initialization, the combined system set a new state-of-the-art for unsupervised transition-based dependency parsing.

It is already interesting that BERT initialization improved the whole system and (in ablation tests) every subset of it, even though neither the system's input data nor its modeling task much resemble the cloze predictions BERT is trained for. But in fact, of the various ablated configurations, those that emphasized dependency graphs most in the inputs and in the output mechanism are the ones where BERT initialization helped the most, strongly suggesting that robust dependency syntax information is implicit in the BERT encodings despite its superficially non-syntactic pretraining task.

Hewitt and Manning (2019) confirm this by direct inspection of ELMo (Peters et al., 2018a, non-attentional contextualized embeddings) and BERT embeddings. They identified projections of the embeddings into low-dimensional spaces, in which distances accurately represent necessary structural properties of a syntactic parse tree: inter-word tree distance and per-word depth. They demonstrated that this outcome was dependent on training for inter-word dependencies, in that it did not occur with a baseline (randomly initialized BiLSTM) that uses word-to-word memory to generate embeddings but is not trained to optimize anything about them.

Reif et al. (2019) replicated their findings, demonstrated that different dependency types characteristically have their own slightly different lengths (with narrow variation), and furthermore showed that other projections of the embeddings represent word sense disambiguation.

Unfortunately, neither paper reported unbounded dependencies separately from the vastly more numerous bounded ones, so the stated results do not prove anything particular about long-distance syntactic dependencies, far less semantic or pragmatic ones.

1.3.1.1.2 Contextualization represents more abstract relationships, less thoroughly Tenney et al. (2019) probed the extent to which limited spans of ELMo, GPT (Radford et al., 2018, attentional encodings), and BERT embeddings contain information from other parts of the sentence, as measured by the spans' ability to inform eight assorted NLP labeling tasks. They found that tasks requiring word-level or syntactic information benefited more from contextualization than those requiring semantic or pragmatic information, understood to mean that the syntactic information was

propagated and encoded into word representations within the span more reliably or clearly.

Up to 79% of ELMo's contextualization benefit could be matched by a control whose word representations were contextualized only locally, which hints at the ratio of local to long-distance information ELMo contains. However, ELMo's advantage over the strictly local context was wider on semantic tasks than syntactic, and (as one would hope) wider when labeling longer dependencies, suggesting that if contextualized representations do not embed long-distance semantic relationships as efficiently as local or syntactic ones, they nevertheless embed them.

Further performance gaps, from the two-layered ELMo encoder up to 12-layered BERT-base and from there to 24-layered BERT-large, were also much bigger for semantic than syntactic tasks, suggesting that an encoder's depth helps to extract/abstract semantics from the details of individual words. The improvement granted by the extra layers of BERT-large was particularly dramatic in the case of the Winograd schema task (Levesque et al., 2012), a pronoun resolution challenge designed to require particularly sensitive use of world knowledge invoked by linguistic context.[8]

Though outscoping is predicted both by syntax and by highly abstract semantic/pragmatic factors, drawing mixed implications from these results, there may be a bright side to the extremely thorough representation of syntax in the embeddings. The linear order of two quantifiers is a scoping factor in its own right, when not redundant with their syntax. In that capacity, it describes their information structure, albeit only to the smallest meaningful degree. Syntactic features used by some previous scope prediction systems, such as a feature identifying a quantificational noun phrase's role use prepositional object, transitive verbal object, or subject, would also identify the use, often information-structurally motivated, of such devices as passivization or dative shift.[9] But descriptions as minimal as these are likely to miss some of the reflections in information structure

[8]Context-sensitivity is ensured by identifying a single-word alternation that changes the preferred antecedent. A well-known example is

(i) The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.

I ended up modeling my approach to the scoping problem on a system for the Winograd challenge; see Section 4.1.

[9]So far as reports of these systems reveal, this feature would, however, give no sign of heavy shift or topicalization.

of other known scoping factors, such as coreference linking (Kuno, 1991) and coherence with other parts of the discourse (Dwivedi, 2013), that a more complete description could capture.

Information structure is, of course, created by choosing syntax so as to manipulate the sequence in which referents, predications, and assertions appear. Since the contextualization process begins by encoding sequence and proceeds to encode syntax, the contextualized embedding of a quantified-over noun potentially contains a better description of its information-structural situation than scope predictors have previously had access to. Whether or not the encoder is able to nail down its information-structural relationships as discourse dependencies, it certainly can represent the syntactic choices that result.[10]

1.3.1.1.3 'Language models are unsupervised multitask learners', to an extent A broad optimism emerges from Radford et al. (2019). They trained GPT-2, an attentional language model of enormous size (by 2019 standards; both by parameter count and by mass of training data), couched diverse NLP tasks in language modeling terms, and announced on the basis of its task performance that 'language models are unsupervised multitask learners'. On closer examination (Appendix A), this is quite true for narrowly linguistic tasks, where the cognitive content of the task is closely akin to small-scale text generation. In such tasks, not only syntax and morphology, but lexical semantics and even some discourse organization (namely semantic prosody) are within its grasp. For more cognitive tasks, and even for linguistic tasks if sufficiently unlike a cloze, a model of stating an answer is not necessarily a good model of finding an answer, and training the latter solely by training the former is, at present, too indirect to be effective.

1.3.1.2 Pessimistic overviews

Two general critiques of language-modeling pretraining apply. First, after a thorough combinatorial study of pretraining and retraining among various tasks, Wang et al. (2019) suggests that language

[10]This is in addition to the fact that the syntactic relationship between two quantifiers, in and of itself in an out-of-the-blue sentence, is already a scope factor underutilized by previous prediction systems other than Manshadi et al. (2013) and perhaps Higgins and Sadock (2003).

modeling alone will not lead to truly general-purpose contextualized embeddings, but that we will need to know more about how different tasks complement each other, how to prevent catastrophic forgetting, and how to maximize the benefit of multitask training.

All of this may be true. Their findings include particularly pervasive indications that the Multi-genre Natural Language Inference (Williams et al., 2017) and Quora Question Pairs (Iyer et al., 2017) tasks rely on information that language modeling does not rely on, that training on language modeling does not provide, and that can benefit other downstream tasks.[11] But by the same token, they found that language modeling was a more beneficial pretraining, on average across target tasks, than any of the supervised tasks were. Together, this is not so much a recommendation against attempting to predict scope on the basis of pure language-modeling pretraining as a recommendation against stopping there.

Second, Chaves (2020) explores the ability of a (non-attentional) contextualized representation, trained on language modeling, to reproduce island effects on the acceptability of filler-gap constructions. Human acceptability judgements here are shaped by a complex array of semantic and pragmatic constraints, but the computational model seems to miss all of these details and learn about the construction only superficially.

Since the expressive capacity of the system that is trained falls above the requirements of human language on the Chomsky (1956) hierarchy, and the training data exceed the amount of language a human learner is exposed to, Chaves concludes that what the model lacks is not found in the training data: The human judgements involve 'rich morphological, syntactic and semantic dependencies which crucially interact with pragmatics and world knowledge' of which the model is ignorant.

This objection strikes right to the heart of the proposal, since the gaps he sees in model knowledge fall squarely within the scoping factors we know humans use. But they are somewhat at odds with what we believe the models do contain. Their world knowledge may have less situated grounding than ours, as discussed in Section 1.1.1, but they are extremely widely read and capable

[11] Based both on these two tasks' effects when used as pre-/intermediate training for others, and on the effects of other pre-/intermediate training on them.

of inferring much; and BERT is trained on multi-sentence inputs, giving it at least the means to learn some of what sentences are used *for*.[12]

The great interest and success in using models like these for sophisticated inference tasks such as SuperGLUE (Wang et al., 2020) demands and reveals, precisely, pragmatic and world knowledge. And Chaves himself has found that some, though not all, attentional language modeling encoders do learn filler-gap much better, though still with some brittleness (Da Costa and Chaves, 2020, discussed below).[13] But it is certainly possible that relevant semantic and pragmatic facts may be lost in the noise or even not captured at all by the pretraining. *A priori* discussion may have taken this point as far as it can go; it may be more useful now to experiment and find out.

1.3.2 Comparison to other tasks

This section discusses specific experimental results on tasks with long-distance dependencies, done with BERT and with other encoders, to fill in the picture of its potential for this application. Overall, it is shown able to noisily capture semantic and pragmatic information about its inputs, although it is stronger on matters of syntax, and the advantages it brought to comparable and related tasks suggest that it brings broad-domain natural-language outscoper identification just into the realm of possibility.

1.3.2.1 Long-distance syntactic dependencies

Long-distance syntactic dependencies include mere morphosyntactic agreement, primarily of interest only because it allows extensive testing of the length of dependencies that an encoder can learn, but also include more abstract and challenging relationships.

[12]The encoder I use, RoBERTa, has different multi-sentence training than original BERT, but both have it. RoBERTa omits BERT's 'next-sentence prediction' training objective but compensates by packing each input window with as many contiguous sentences as possible.

[13]BERT performed the best of those tested. GPT-2 did well also. Two others, Transformer-XL and XLNet, did worse than the previously studied non-attentional models on the filler-gap task.

1.3.2.1.1 Morphosyntax Da Costa and Chaves (2020) probed the learning of filler-gap dependencies in LSTM- and Transformer-based neural models, using both wh-extraction and clefting across 1–4 clause boundaries. Stimuli had gaps in subject positions, such that their verbs had to agree in number with the filler. Given the sentence up to the verb as context, if the dependency has been learned, surprisal should be lower for the form of the verb with correct number agreement than without. A second set of stimuli probed models' ability to await a gap after encountering a filler, without any agreement requirement. Here the gap was the object of a preposition, and the crucial measure was surprisal of the post-gap word vs. surprisal of an ungrammatical resumptive pronoun.

Transformer-XL had lower surprisal for the agreeing verb only across a single clause boundary, and got the resumptive pronoun test exactly backward. XLNet was similarly poor.

BERT got correct agreement across four boundaries for it-clefts, and across three for wh-questions, but only by making use of the rightward context of the verb; in a variant of the resumptive-pronoun stimuli that made this following context unhelpful, surprisals were precisely the opposite of what they should have been, even across just one clause boundary: low surprisal for a fillerless gap or a gapless filler, high surprisal for both or neither. Da Costa and Chaves conclude that BERT learns filler-gap constructions, but not robustly.

GPT-2 outperformed BERT on the basic tests, but also broke down in certain constructions when following context was unhelpful. On tests they passed, both of these models showed clearer distinctions in surprisal at lower levels of embedding, but it would not surprise me to see human subjects also uncertain when required to produce number agreement across four clause boundaries.

Goldberg (2019) finds that a BERT-based system is able to correctly predict reflexive anaphora, though doing so across clause boundaries was the least successful of all his anaphora and agreement tasks.

Lin et al. (2019) probed BERT's learning of subject-verb agreement and reflexive anaphora by examining attention weights on the correct subject or antecedent, when processing the verb or anaphor, in the presence of theoretically and psycholinguistically motivated distractors. Decreased attention to the true trigger increases a statistic termed 'confusion', which was calculated at each

layer. Distractors may or may not share the true target's agreement features, and may or may not be in appropriate syntactic positions (same clause versus relative clause for subject-verb, c-command for anaphora).

In subject-verb agreement, BERT distributes its attention widely even in unambiguous cases, but confusion is higher when distractors are in the same clause and when they match the subject's number. Thus it is attending, albeit weakly, to the proper linguistic features.

In reflexive anaphora, all distractors raise confusion, including those that are featurally incorrect and those that do not c-command, but again, confusion is higher for gender-matched distractors and slightly higher for those that c-command. Once again, this is the correct signal, but in heavy noise. Humans' 'sharp sensitivity to hierarchical structure' is not reproduced.

For both tasks, within the encoder the layer-wise trend is for lower confusion in higher layers, suggestive of the morphosyntactically relevant facts being progressively abstracted away from details of the input.

1.3.2.1.2 Toward pragmatics, toward scope Suites of tasks illustrated that as task demands shifted further from surface patterns and toward semantics and pragmatics, success was no longer so easy as with morphosyntax. Acceptability of negative polarity items, which depends on semantic scope of negations rather than on syntax as such (Ladusaw, 1979), was one of the most difficult of these.

Warstadt et al. (2019b) prepared and validated a broad-domain corpus of minimal sentence pairs (in which a single change toggles grammatical acceptability), each characterizing a single phenomenon. They evaluated GPT-2 and Transformer-XL attention-based language models on their ability to assign higher probability to the grammatical sentence in each pair. No task-specific supervision was provided.

These models also coped well with morphological agreement. They fell furthest short of human performance on islands, negative polarity items (GPT-2 accuracy 78.9% vs. human 88.1%), and quantifiers (GPT-2 accuracy 71.3% vs. human 86.8%). The latter two involve pragmatic reasoning

and semantic distinctions increasingly distant from syntax, precisely the sort of task I propose, and correct use of NPIs requires inferring a semantic scope.

Warstadt and Bowman (2019) report a similar study with BERT and GPT, using the fine- and coarse-grained subsets of CoLA (Warstadt et al., 2019c) as a survey of syntactic phenomena such as adjunction, complement clauses, extraction islands, or ellipsis.

Interrogatives were associated with notably lower performance, which the authors ascribe to long-distance dependencies created by *wh*-extraction. With occasional exceptions, lower performance was also associated with movements for focus/information structure, dislocations (such as heavy shift), coordination,[14] subordination (such as relative clauses), conditionals—that is, constructions often associated with filler-gap and/or anaphoric dependencies—and interesting determiners such as quantifiers.

BERT and GPT greatly outperformed a baseline without attention in cases involving unbounded dependencies, particularly reflexive pronouns and dislocations, presumably because long-distance comparisons are the very essence of attentional decoding. However, they did notably worse than baseline on sentences involving negating auxiliaries and/or negative polarity items, so they may not be capturing scopal relationships between words as effectively as purely syntactic ones.

Warstadt et al. (2019a) follow up on negative polarity items and BERT. They find that BERT (with no fine-tuning) does in fact represent all of the features of NPI licensing, but its ability to demonstrate this knowledge varies dramatically with experimental method, and its knowledge of negation's scope is weaker.

1.3.2.1.3 Textual scope McKenna (2019) trained a system to predict the textual scope of negation, i.e. the span of words considered to be negated, rather than a containing term of predicate logic. Strangely, he got near-state-of-the-art performance from a tree neural network that abstracted

[14] Since CoLA items are selected from linguistics publications, they heavily represent unusual uses of 'and', among them non-constituent coordination, ellipsis, and causative readings.

a sentence to just its syntax, and performance was slightly degraded by adding and fine-tuning contextualized word embeddings, whether BERT or learned from scratch. Using BERT embeddings without fine-tuning degraded performance substantially.

Whatever the reason for this aggressively anti-semantic outcome, I do not expect it to generalize to quantifier scoping. From experience as an annotator, scoping judgements heavily involve world knowledge.

Sergeeva et al. (2019) predicted the textual scopes of supposition and negation from various contextualized word embeddings. Fine-tuned BERT outperformed ELMo, which in turn outperformed pretrained BERT. Adding explicit syntax features to fine-tuned BERT did not improve performance, whereas it did with the other two embeddings.

I take these results to mean that syntactic knowledge useful for this task, but not for the pretraining word prediction tasks, was acquired in the course of language modeling, but fine-tuning was necessary to get it out of the encoder.

Both studies show a surprisingly tight connection between syntax and the textual scopes they study. This does put it in question whether they are really analogous to quantifier scoping, where nonsyntactic factors are highly relevant.

1.3.2.2 Various other relevant tasks

We now pivot to an array of more semantic tasks that may be similar to outscoper identification in nature or difficulty, including semantic parsing, entailment recognition, and some context-sensitive inference tasks.

Pütz and Glocker (2019) adapted a state-of-the-art transition-based semantic parser by replacing its custom recurrent encoder (for input words and output graph states) with a small convolutional network atop ELMo embeddings. When supplemented by silver training data from the baseline systems, this approach achieved competitive results. Semantic parsing may not involve particularly long-distance dependencies, but as a meaning-oriented task I judge it at least as relevant to my scoping challenge as filler-gap was.

Wang et al. (2018) prepared a diagnostic set of entailment recognition tasks, labeled as to particular linguistic/semantic phenomena they depended on, for the purpose of comparing the strength of various contextual encoders (some attentional, all pre-BERT). As a guide to the difficulty of the various phenomena, we can compare reported performance on labeled subsets to average performance across the whole diagnostic set.

Among the fine-grained subsets, models handled the Universal Quantification subset well, but deducing entailment from universal quantification does not necessarily require good scope prediction. In the absence of a second, different quantifier, most of these entailment tasks amount to recognizing an entity in one sentence as belonging to the restrictor set of the universal in the other sentence. For comparison, models performed only half as well as whole-set on the Downward Monotone subset,[15] to which scoping has the same partial relevance. Although quantification is involved, these subsets are probably not very informative for judging the prospects of scope prediction.

The coreference subset was also handled well, which is mildly encouraging since both coreference and scope disambiguation are pragmatic judgements about the relationship between two nouns or pronouns, but coreference is certainly the eaiser task. However, worst-performing of all the fine-grained subsets reported was Restrictivity, a phenomenon tightly intertwined with scoping and quantification. So the prognosis thus far leans negative.

Of the coarse-grained subsets, only Logic had a preponderance of phenomena relevant or similar to quantifier scope prediction. With a few exceptions, models' performance on this subset was 70–80% of whole-set average. The noteworthy exceptions are language models trained solely as encoders for entailment detection and supplemented by CoVe (attentional machine translation-based; McCann et al., 2017) contextualized word representations. These were relatively stronger in the Logic subset than others, though still not as strong as on other, less challenging parts of the task set.

Impressionistically, the Wang et al. (2018) diagnostic results suggest that contextualized representations had then become barely sophisticated enough to make scope prediction possible. And, as

[15] Not all downward monotone terms are quantifiers, but the others (negations) are still scopal.

noted, BERT made clear advances over this state of affairs in tasks of all sorts.

Finally, out of the tasks that Radford et al. (2019) attempted to reduce to language modeling, three I find particularly relevant are the Winograd Schema Test (Levesque et al., 2012), LAMBADA (Paperno et al., 2016), and Natural Questions (Kwiatkowski et al., 2019). As mentioned in Section 1.3.1.1.2, the Winograd Schema Test requires the calculation of coreference that is highly sensitive to the semantic context, which is a reasonably close parallel to scope-as-dependency.

LAMBADA is a sentence-final cloze task, selected to be impossible from the sentence alone, but possible with the preceding 50 tokens as context. Unlike in coreference or scoping, the correct answer does not necessarily have a literal antecedent in the materials of the question; but as in scoping, it can depend on the semantics of the situation described and/or the direction the discourse takes, and may require assembling multiple clues from across the context.

Natural Questions requires extracting a word or phrase from encyclopedic text, in answer to questions for which that text may not have been designed, possibly with the use of knowledge not found in the text, and often with reference to the relationships among entities the text describes. Again, that is a reasonably good description of identifying outscopers in our corpus materials.

To summarize their findings, the GPT attentional language model performed not too badly on the Winograd Schemas and LAMBADA, relative to the difficulty of the task, but underperformed dramatically on Natural Questions. This may in part be an artifact of trying to stuff all the tasks into a language modeling framework as directly as possible, since Alberti et al. (2019) did succeed in getting BERT to answer Natural Questions with some supervised training and an ensemble-like approach to generating the answer.

1.4 Prospects and plans

To summarize, BERT and similar models encode numerous intra-text dependencies, but they are strongest on the syntactic and local ones. Plausibly, this is because their pretraining to model language supplies a better, less ambiguous learning signal for syntax than for semantics, and many

more examples of local dependencies than long-distance. Optimizing for their cloze training task implicitly optimizes for a lot of syntactic analysis and a certain amount of world knowledge. This, I believe, underwrites most of their success on assorted NLP tasks.

Using very high-dimensional representations leaves room for training to pick up subtler connections among words and/or among the concepts they represent, information that is relevant to semantic and pragmatic tasks such as outscoper identification. I believe this is the auxiliary factor that extends their success to tasks that are more reasoning-oriented. But two obstacles intervene. The first obstacle is that these signals are surrounded by a lot more noise. They are not so constantly in evidence as facts of constituency or semantic prosody are. The second obstacle is that these subtler factors just play vastly less of a role in the selection of single words than local or even long-distance syntax. Although no blatant failures recommend frank pessimism about applying a BERT-like to tasks like these, it is unlikely to be an easy success.

1.4.1 Encode-and-classify with RoBERTa

RoBERTa is the result of searching for ways to improve BERT's pretraining, and was state-of-the-art for many tasks (Liu et al., 2019) when this project reached the point of selecting an encoder. For the job of determining whether these embeddings have anything to offer for scoping, it was therefore a logical choice. Reasons for selecting it are discussed more fully in Section 4.3.1.

The encodings of two quantified-over nouns would then be fed to a classifier to categorize the relationship between them, as in the Manshadi et al. (2011) evaluation, as direct, inverse, or non-scopal. Fine-tuning the encoder with task data, as opposed to using it as a feature extractor and only training the classifier, usually improves task performance, and so was adopted as part of the plan.

Since the encodings of the nouns could reasonably be expected to include the full range of scoping factors, this design potentially results (if not necessarily immediately) in a substantially more general scoping system than has previously been built.

1.5 Previous quantifier scope disambiguation systems

The semantic task of determining all possible scopal readings of a sentence can be addressed with compositional rules, and the task of identifying the weakest readings can be done algorithmically (Koller and Thater, 2010). However, the pragmatic task of identifying the *preferred* scoping has not been solved in the general case.[16]

1.5.1 Descriptive efforts

We here group several scope-prediction attempts that rely on hand-written heuristics devised by inspecting scoping judgements on a small number of artificial example sentences. Empirical measures of prediction quality are typically not reported.

As part of the Lunar Sciences Natural Language Information System (Woods et al., 1972), VanLehn (1978) studied syntactic influences on quantifier scoping and developed three theories of it. The first theory introduced syntactic structure sufficient to reduce every quantifier scope ambiguity to a syntactic ambiguity. The second drew an analogy between anaphora and universal-over-existential scoping, on the basis of syntactic structures that disfavor both. The third postulated a lexical iterability property of each predicate, and a syntactic 'iteration phrase' that is capable of iterating over multiple variables in parallel but that is only felicitous when dominating a highly iterable predicate. By way of measuring empirical performance, VanLehn reports judgements from multiple informants on the sentences that illustrate how these theories apply. However, he concedes that 'none of the three theories predicts the data with an accuracy that demands conviction' (VanLehn, 1978, 18).

TEAM (the Transportable English database Access Mechanism) used heuristics of syntax, word order, and quantifier lexical realization to score scopings of natural-language database queries (Martin et al., 1983).

[16] Manshadi et al. (2013) may be quite strong, but limitations of the data confound the issue. Highly successful algorithms are available for certain special cases (Evang and Bos, 2013; Schuler and Wheeler, 2014).

	Precision
Baseline	61.2%
Naïve Bayes	72.6%
Maximum entropy	73.5%
Perceptron	77.0%
Re-annotation	76.3%

Table 1.1: Higgins & Sadock (2003) empirical accuracy.

The SRI Core Language Engine employed a Cooper store (Cooper, 1983) and imposed six rules and eight preferences on withdrawing quantifiers from the store to impose scoping (Moran, 1988). The engine's compliance with the rules and preferences was tested, but the empirical validity of the rules and preferences was not.

Another, similar set of scoping heuristics was described and implemented by Hurum (1988).

1.5.2 Higgins and Sadock (2003) and WSJ

Higgins and Sadock used machine learning to identify features that predict quantifier scoping in sentences that contain precisely two quantifiers other than the determiners *a, an, the,* extracted from the WSJ portion of the Penn Treebank.

Their single-layer perceptron learned that truth-conditional equivalence of the two possible scopings (i.e. scopal non-interaction) was predicted by a comma, colon, or conjunct intervening between quantifiers, or by the second quantifier being *all.* Inverse scoping was favored when the second quantifier was *each* or *most.* In-situ scoping was favored when the first quantifier c-commanded the second, or when a clause boundary intervened. Maximum-entropy and naïve Bayes classifiers produced similar results to the perceptron. Empirical accuracy is shown in Table 1.1, with two standards of comparison: a baseline system that always predicted non-interaction, and an independent re-coding of the sentences by a second annotator.

Of the traditional four scoping factors summarized by AnderBois et al. (2012), this project made use of three: Quantifier lexis, quantifiers' linear order, and syntax. Its coverage of syntactic features was extensive, not outdone or even matched until QuanText (Manshadi and Allen, 2011; Manshadi

et al., 2013). Like all of the earlier work reported here, it did not employ the factors related to connected texts. Apart from this, its major weakness is the very small diversity of quantifiers it models.

1.5.3 Andrew and MacCartney (2004)

Andrew and MacCartney predicted quantifier scoping in two-quantifier sentences extracted or adapted from Law School Admissions Test (LSAT) logic puzzles, using naïve Bayes, logistic regression, and support vector machine classification. These systems performed two binary classifications.

For sentences whose truth conditions depend on quantifier scoping, they predicted whether the correct scoping was in-situ or inverse (the *scope direction* task). As a baseline, always predicting in-situ produced 82.9% correct predictions; the best of their classifiers raised this to 94.3%.

Adding to these a second class of sentences, those in which quantifier scoping is irrelevant to truth conditions, they predicted which class each sentence belonged to (the *scope interaction detection* task).[17]

As a baseline, always predicting an interaction produced 76.1% correct predictions, a standard which none of their classifiers beat and which only one classifier matched.

Their results on the scope direction task may make it appear that scope prediction is largely a solved problem. However, as the baseline score reveals, this is partly an artifact of the great uniformity of their data. All sentences were edited down to exactly two quantifiers, and logic puzzles are constructed to be unambiguous or easy to disambiguate.

Moreover, logic puzzles are constructed to be solvable given only what is stated in the problem; in other words, to specifically avoid any contribution from world knowledge. Thus, although the project accepted a wide variety of quantifiers for prediction, its other linguistic diversity was limited and the breadth or narrowness of its subject matter was deliberately rendered irrelevant.

[17] Note that the two tasks were tested independently; that is, the scope prediction classifiers ran on the entire test set, not just the sentences in which another classifier detected the scopal interaction.

The predictor relied on the factors of lexis, linear order, and two syntactic features per quantifier: whether it fell within the syntactic scope of a negation, and whether it fell within a conjunction whose other conjunct contained the other quantifier.

1.5.4 Srinivasan and Yates (2009)

Srinivasan and Yates built a classifier to predict preferred and plausible scoping in predicates whose two arguments are quantified by *a/an* and *every,* using world knowledge heuristically extracted from unlabeled text. The heuristic takes numeric quantifiers in the training text as indications of the typical size of sets of entities; in particular, a training sentence like 'The grand jury returned three indictments' is taken to suggest that three is a typical size both for sets of indictments and for sets of things that are returned. On such grounds the classifier predicts the size of the restrictor sets for *a/an* and *every* in its test items. The relevant intuition is that for *a/an,* a small or singleton restrictor set correlates with wide scope, whereas a restrictor set about as large as that of *every* correlates with falling under the scope of *every.*

Though this study treats only two quantifier types and only two possible scopings per item, the data (taken from the Web1Tgram Corpus) are much more naturalistic (i.e. unruly) than those of Andrew and MacCartney (2004), and the task performance correspondingly lower: A baseline always predicting that in-situ scope is preferred had an accuracy of only 53%, versus 74% for the trained classifier. A baseline always predicting that inverse scope is plausible had an accuracy of 67%, versus 73% for the trained classifier. I am impressed that only a single, rough-hewn form of world knowledge was able to produce such performance gains.

The other scoping factors were absent, and in addition to extremely limited quantifier diversity, other linguistic diversity is absent, since the data consist of Minimum Recursion Semantics (Copestake et al., 2005), and not of natural language at all.

1.5.5 Dinesh et al. (2011)

Dinesh et al. (2011) annotate the scoping of quantifiers, modals, and other operators in 195 sentences

of FDA regulations (average 30 words each). They predict each operator's most likely outscopers with a maximum-entropy classifier, as part of a project in automatically verifying compliance with regulations (Dinesh, 2010).

Under the most generous of their metrics, scoring the prediction as accurate or inaccurate for each pair of operators, their classifier achieved an F-score of 90.6%. They themselves describe this score as "inflated by inclusion of reflexive pairs" (correctly classifying the scope of an operator relative to itself) and conclude that "it is better to consider the relative improvement in F-score over the [in-situ] baseline," which was only 36.6%. They also offer a less inflated scoring measure, on which the accuracy of their best model is 69.4%.

In a scope prediction task confined to quantificational determiners, but also confined to predicting *de re* or *de dicto* scoping, their classifier reached an accuracy of 81.2%.

Regardless of the exact metric, these results suggest there is room for improvement in scope prediction. Moreover, like logic puzzles, regulations as a genre are meant to be unambiguous and are allowed considerable verbosity toward that end. Most genres must trade unambiguousness for brevity, and in these the prediction problem will be correspondingly harder.

Of the classic scoping factors, quantifier lexis is partly reflected by a feature classifying the quantificational determiner in a partly semantic, partly lexical scheme. A linear order feature is used: the quantifier's position relative to its clause's main verb. Since this project predicts only *de re/de dicto,* which in scopal terms amounts to the quantifier's scope relative to an implicit 'postcondition' operator, this feature is more or less analogous to the linear order of two quantifiers used elsewhere. The abstraction used in the prediction, called a Processed Parse Tree, is flatter than a full-blown syntactic analysis, but contains structures for e.g. prepositionally modified noun phrases. Some use of world knowledge may inhere in working with texts consisting exclusively of food and drug regulations, though this is not clear. In any case, a scope predictor so trained may not generalize to other domains or other, less exact linguistic registers.

1.5.6 AnderBois et al. (2012)

AnderBois et al. (2012) return to the LSAT logic puzzle genre and improve data quality by having multiple annotators code each item. They cover only non-cumulative two-quantifier sentences with at least one quantifier as subject or direct object (the other quantifier may perform some other function).

They model four predictors of the subject or object quantifier's scope:[18] its grammatical function, the two quantifiers' lexical realizations, and the relative order of the two quantifiers. All four independently predict scoping, with similar effect sizes.

AnderBois et al. (2012) write in the corpus linguistics tradition, rather than computational linguistics, and so do not provide an accuracy score against a test set, but their statistical model accounts for 84.7% of the variability in their dataset, in spite of considering the syntactic scoping factor only minimally. The same caveats about the nature of the data apply here as for Andrew and MacCartney (2004): It is linguistically homogeneous and prevents world knowledge from being useful, which exaggerates the model's effectiveness at capturing the full natural phenomenon.

1.5.7 Manshadi et al. (2013)

QuanText, by Manshadi et al., consists of 500 imperative sentences similar to Example (2), giving instructions for manipulating text files.

(2) Print every line of the file that starts with a digit followed by punctuation.

This corpus forms the basis of the first attempt to statistically predict quantifier scope over materials with such complexities as more than two quantifiers. (Manshadi and Allen, 2011). A support vector machine was given a large and sophisticated set of lexical and grammatical features and achieved complete recall of scoping in 72% of test sentences (Manshadi et al., 2013), a figure made

[18] In sentences whose two quantifiers were subject and direct object, one of the two was taken at random as the quantifier of interest.

more impressive by comparing it to QuanText's inter-annotator agreement of 75% (Manshadi et al., 2012). However, this result must be read cautiously.

The principal objects of QuanText's domain—characters, words, lines, and files—are overwhelmingly in part–whole relationships, so that a very simple scoping heuristic based on these head words has the same 72% per-sentence complete recall accuracy as the support vector machine (Schuler and Wheeler, 2014). This confirms that world knowledge contributes usefully to quantifier scope disambiguation, but probing its full contribution will require predicting scopes in broader subject matter.

This project's syntactic predictors are a superset of those from Higgins and Sadock (2003), using both phrase-structure and dependency representations. The phrase-structural features are rich enough to imply quantifier linear order as well. Quantifier lexis is used. By encompassing sentences with more than two quantifiers, this project includes previously neglected linguistic diversity—offset somewhat by the fact that all of its sentences are imperatives. Its quantifier diversity is extremely good. It is primarily the issue of its single, very regular domain that brings some dissatisfaction (and, as always, the single-sentence limitation that misses discourse-related factors).

1.5.8 Tsiolis (2020)

Tsiolis (2020) reports two unsuccessful applications of pretrained attentional language models to quantifier scope disambiguation. The first attempt formulates scoping as a natural language inference problem, then uses BERT fine-tuned on MNLI (Williams et al., 2017). The original sentence is given as premise, then BERT is asked whether it entails a hypothesis consistent with one or the other scopal reading. Unfortunately, the system predicts entailment for every hypothesis.

The second attempt paraphrases a subset of the Higgins and Sadock (2003) data into forms with less scopal ambiguity, then scores the paraphrases' probability, operationalized by their GPT-2 perplexity. Unfortunately, this method fared much worse than the baseline of always predicting direct scope.

Tsiolis also reports personal communications from Justyna Grudzińska, Aleksander Wawer, and Marek Zawadowski about two unpublished systems. One, with BERT, is reported to have been unsuccessful. The other, with a different sentence encoder, is reported to have been successful within the AnderBois et al. (2012) logic puzzle data, taking advantage of predictable scopal effects of certain prepositions. This apparently relates to the theoretical work in Grudzińska and Zawadowski (2020).

Chapter 2

Building the scope-annotated corpus

2.1 Prior scope corpora

The semantic task of determining all possible scopal readings of a sentence can be addressed with compositional rules, and the task of identifying the weakest readings can be done algorithmically (Koller and Thater, 2010). The pragmatic task of identifying the *preferred* scoping is unsolved in the general case, although successful algorithms are available in certain narrowly limited cases (Evang and Bos, 2013; Schuler and Wheeler, 2014).

To build a general-purpose model of human scoping preferences calls for training data not so narrowly limited in syntax or subject matter. Indeed, it calls for data covering the full range of cues that correlate with human scoping judgments: text coherence (Dwivedi, 2013), linear order of scope-bearers, syntactic structure, choice of lexis, and use of knowledge about the world (AnderBois et al., 2012). The world knowledge humans rely on may be partly expressed explicitly in the text being interpreted, and partly presumed background knowledge; the data should realistically reflect this. Table 2.1 uses these criteria to review previously collected scope data.

2.1.1 VanLehn (1978)

VanLehn studies syntactic influences on quantifier scoping, as an outgrowth of an effort to improve scope disambiguation in the Lunar Sciences Natural Language Information System (Woods et al., 1972). He reports that 'well over 1500 [quantifier scope] judgments [...] and hundreds of pages

	Size	Genre	Rich	Broad	Order	Lexis	Parse	World	Text
VanLehn	> 1500 pairs	?	yes	?	yes	yes	yes	?	?
Higgins & Sadock	893 pairs	news	no	yes	yes	yes	yes	yes	yes
Andrew & MacCartney	305 pairs	logic puzzle	yes	?	yes	yes	yes	no	no
Srinivasan & Yates	92 pairs	artificial	no	yes	no	no	no	yes	no
Dinesh et al.	195 sent.	regulatory	yes	no	yes	yes	yes	yes	yes
AnderBois et al.	358 pairs	logic puzzle	yes	?	yes	yes	yes	no	no
Manshadi et al.	500 sent.	instructions	yes	no	yes	yes	yes	no	no
Evang & Bos	456 pairs	(multiple)	no	yes	yes	yes	no	yes	yes
Current work	2000 sent.	encyclopedic	yes	yes	yes	yes	yes	yes	yes

Table 2.1: Quantifier scope corpora and criteria for their use as training data. The criterion *Rich* refers to the density of multi-quantifier sentences in the genre. *Broad* refers to subject-matter coverage. *Order* is quantifiers' in-sentence sequence; *Lexis* is the words expressing them; *Parse* is their use in varied syntactic environments; *World,* their use where general knowledge is presumed; and *Text,* use in connected discourses.

of natural text' were collected in this effort, but with 'inconclusive' and 'remote' prospects for improving the system. The later fate of this system, and thus of the scope data, is not known to us.[1]

2.1.2 Higgins and Sadock (2003) and WSJ

Higgins and Sadock also studied quantifier scoping. They identified 893 sentences having two quantified expressions within the WSJ subset of the Penn Treebank (out of 41,191 sentences total), and annotated each as having in-situ scoping, inverse scoping, or no scopal interaction. The *no interaction* category proved to be dominant, with 545 members (61%).

Machine learning on this corpus identified some features conducive to each outcome. For example, in their single-layer perceptron, scopal non-interaction was predicted by a comma, colon, or conjunct intervening between quantifiers, or by the second quantifier being *all.* Inverse scoping was favored when the second quantifier was *each* or *most.* In-situ scoping was favored when the first quantifier c-commanded the second, or when a clause boundary intervened. Maximum-entropy and naïve Bayes classifiers produced similar results.

A second, independent coder agreed with the reference coding on 76% of sentences, at a Cohen κ of 0.52. Unsurprisingly, the sentences on which coders agreed were also predicted more accurately by the classifiers. Again taking the perceptron as an example, it agreed with the annotators on 83% of these, versus 77% on the test set generally.

This corpus comes near to meeting the modeling requirements. It provides linear order, syntax, lexis, and presumptions of world knowledge. The coherent text from which the sentences were drawn is available. We have chosen not to reuse it for two reasons.

The first reason is the low overall density of quantificational noun phrases in this text genre, reflecting the fact that news writing is about particulars. Our interest lies in explanatory text, which concerns generalities and (in our experience) quantifies much more often.

The second reason is the phenomena deliberately avoided. Higgins and Sadock did not treat the determiners *a, an, the* as quantifiers of interest, which in their words 'avoids the problem of generics

[1] It is frequently cited in later literature, but the VanLehn paper seems to be the last allusion to active work on it.

and the complexities of assigning scope readings to definite descriptions.' Limiting the scope of the problem is a legitimate choice, of course, but both phenomena are prominent in explanations such as Example (1).

(1) Bus runs must keep to a timetable. The driver may need to hurry or delay to stay on schedule.

In this example, the generic *bus runs* states a generalization, and the definite *the driver* is subordinated to it in order to elaborate on the statement.[2]

2.1.3 OntoNotes

The Penn Treebank WSJ is one of the foundations of OntoNotes (Pradhan et al., 2007), and for interoperability of resources it is worth considering whether to build a scope corpus as an OntoNotes overlay. The requirements of text coherence, order, syntax, lexis, and presumption of world knowledge remain met as before. Unfortunately, the text's low quantifier density remains a problem, one that carries over to the broadcast news portion due to the similarity of genre.

The other sections of OntoNotes are informal writing and conversation, which bring with them the additional complexities of social identity, turn-taking cues, incomplete sentences, imprecise language due to time pressure, and the fact that multiple participants may not even have the same understanding of the situation they discuss. We wished to avoid these complexities, at least for the time being, and so declined to attempt an annotation layer for OntoNotes.

[2]We construe generics as quantificational, as argued by Leslie (2015) and as assumed by the 'implicit universal' of Manshadi et al. (2013). But their quantification is more like that of *many* than of *some*, in that its meaning depends on the discourse context. We have paraphrased this meaning as 'enough that you should know about it, given the question (expressly or implicitly) under discussion,' taking 'question under discussion' in the sense of Roberts (2012), or less formally as 'more than you might think' (with the understanding that, until the subject came up, you might not have thought about it at all).

2.1.4 Andrew and MacCartney (2004)

Andrew and MacCartney found a quantifier-rich genre in logic puzzles from the Law School Admissions Test (LSAT), and were able to acquire 305 two-quantifier sentences either by extracting them directly or by editing them down from more complex sentences. Their inventory of quantifiers does include *a, an* but still not *the,* nor generics expressed by bare plurals and bare mass nouns.

They do not mention inter-annotator agreement, but this is reasonable enough, since these texts are designed to minimize ambiguity in order to have a single right answer. This has some strange effects; the *no interaction* class that predominated in the WSJ comprises only about 20% of their data, whereas 70% of their sentences were scoped in-situ.

The puzzle genre suits their purpose well, since it reduces the importance of world knowledge and discourse context for interpretation. But these are factors we wish to model, so these somewhat artificial texts suit us poorly despite providing order, grammar, and lexis.

2.1.5 Srinivasan and Yates (2009)

Srinivasan and Yates labeled 92 semi-synthetic quantifier scope disambiguation problems, as test data for an experiment in automatically extracting world knowledge from unlabeled text.

From the Web1Tgram corpus they semi-automatically extracted 128 predicates of two arguments, then identified 46 n-grams in which both arguments to such a predicate were named classes (or could be edited to supply a named class). From each of these 46, they constructed two quantifier scope disambiguation puzzles, each quantifying one class with *a* and the other with *every,* and labeled the preferred scoping as direct or inverse.

Their classifier then predicted preferred and plausible scoping based on the numbers that co-occur with each named class (as in ‘billions of stars’ or ‘fifty million Frenchmen’) and each predicate’s argument position (as in ‘bought a dozen eggs’), taking this as a readily available source of world knowledge on the typical size of such sets.

These are unquestionably broad-domain data, and textual co-occurrence is known to capture world knowledge, as demonstrated by prediction- and count-based vocabulary embeddings (Mikolov et al., 2013; Pennington et al., 2014). But by the same token, these are fairly artificial data.

The labeled quantified expressions are not themselves naturally occurring. They have no discourse context, an absolute requirement for us. They all have precisely two scope-bearers and all use the same two quantifiers. Only two scopal relationships were possible; since a predication must fall within the scopes that bind its arguments, one quantifier always outscoped the other, and since the two scopings of these quantifiers are not truth-conditionally equivalent, there was no opportunity for a *no interaction* label. This limited range of possible scopings is not representative of explanatory generalizations as a whole.

Finally, though the training data for the classifier were natural-language n-grams, these test items were already in an MRS-like logical form (Copestake et al., 2005). Linear order is destroyed, semantic roles are abstracted away from their syntactic realization, and predicates are abstracted away from their lexical realization. The presumption of world knowledge is the only criterion of ours that this corpus does meet.

2.1.6 Dinesh et al. (2011)

Dinesh et al. (2011) annotate the scoping of quantifiers and other operators in 195 sentences of FDA regulations (average 30 words each). This is explanatory text, specially enriched in modals. They predict each operator's most likely outscoper with a maximum-entropy classifier, as part of a project in automatically verifying compliance with regulations (Dinesh, 2010).

For our purposes, three causes for concern are the narrow subject matter (limiting the variety of lexis), the deeply specialized world knowledge that it presumes, and the small size of the corpus. However, text coherence, order, and syntax are intact, and there may be advantages to regulations' position at the interface of exact legal reasoning with complex real events. Regulations are meant to convey meaning more exactly than many other explanations do, which suggests they would provide

a clear training signal. Nevertheless, regulations are forced to grapple with complexity; they must address a certain subject matter, whereas an LSAT problem (for example) can be written about whichever domain allows for the least interpretive ambiguity. For these reasons, the Dinesh et al. (2011) corpus might at a future date prove to be valuable supplementary training data.

2.1.7 AnderBois et al. (2012)

AnderBois et al. (2012) return to the LSAT logic puzzle genre and improve data quality by having multiple annotators code each item. They report 358 observations within the scope of their investigation—non-cumulative two-quantifier sentences with at least one quantified noun phrase as subject or direct object (the other quantified noun phrase may perform some other function). They imply the existence of other annotated data in the corpus, though without mentioning its quantity.

They model four predictors of the subject or object quantifier's scope:[3] its grammatical function, the two quantifiers' lexical realizations, and the relative order of the two quantifiers. All four independently predict scoping, with similar effect sizes.

As before, we worry that the genre is unnatural. Moreover, it purposely limits the role of presumed world knowledge.

2.1.8 Manshadi et al. (2013)

QuanText, by Manshadi et al., is to our knowledge the most thoroughly developed corpus of scope annotations. It consists of 500 imperative sentences, giving instructions for manipulating text files.

Sentences were derived from tutorials, help documents, a survey of computer users, and crowdsourced descriptions of data manipulation demonstrations (Manshadi et al., 2011).

QuanText is the first scope corpus to consider all NP chunks as candidate scope-bearers, including indefinites, definite descriptions, and generics; the first to embrace the complexities added by negation, modals, or sentential adverbs; and the basis of the first attempt to statistically predict

[3] In sentences whose two quantifiers were subject and direct object, one of the two was taken at random as the quantifier of interest.

quantifier scope over such complex materials (Manshadi and Allen, 2011). This more comprehensive coverage aligns with our goals.

Another point in favor is that QuanText sentences, like explanations, routinely contain three or more scope-bearers. This necessitates a more complex annotation scheme than the ternary classification that sufficed for previous projects, and as we shall see, this incurred some problems in the methodology for comparing annotations with one another or with machine predictions. Nevertheless, the closer resemblance to our genre of interest is welcome.[4]

The sole disadvantages of QuanText for our purposes are its narrow subject matter and its lack of any connected, multi-sentence discourse. Subject matter in particular turns out to be a severe disadvantage.

The principal objects of the domain—characters, words, lines, and files—are overwhelmingly in part–whole relationships, so that a very simple scoping heuristic based on these head words has the same per-sentence complete recall accuracy (72%) as a support vector machine using a large and sophisticated set of both lexical and grammatical features (Schuler and Wheeler, 2014; Manshadi et al., 2013), and both are comparable to QuanText's 75% inter-annotator agreement (Manshadi et al., 2012). This confirms that world knowledge can be an effective clue for quantifier scope disambiguation, but neither the small amount of world knowledge that explains so much of QuanText, nor the crude representation of it as merely the identities of head words, promises to generalize well to a broader domain. The data are just too regular to train on.

Finally, the QuanText sentences have been edited to be understandable out of the blue. This prevents any investigation of text coherence or other discourse influences on quantifier scope disambiguation. But natural language is rarely out of the blue, and a cognitive model should account for this.

[4]Not every genre shares this tendency. In WSJ, for example, Higgins and Sadock (2003) found a mere 61 sentences with three quantifiers from their list, and 12 sentences with four.

2.1.9 Evang and Bos (2013)

Evang and Bos extracted from the Groningen Meaning Bank (Basile et al., 2012) all occurrences of PP modifiers with one of *every, each, all* quantifying either the modificand or the prepositional object. Aside from fixed expressions and other exceptional cases, 456 were found and annotated. Because the syntactic environment is so specific, a binary attribute on the preposition suffices to capture the scoping information.

This corpus thus features text coherence, order, syntax, and a particular emphasis on the lexical realization of universal quantifiers. Its subject matter is broad, allowing for modeling of world knowledge effects.

Evang and Bos are optimistic about the potential for automatically generating a near–gold-standard layer of binary scope attribute tags for the GMB, using their hand annotations as training data. But we do not expect these annotations to adequately train scope prediction in a broad-coverage integrated system. Evang and Bos acknowledge that the selected phrases and the purely binary annotation are not adequate to model the semantic/pragmatic behavior of specific indefinites, scopal interactions between quantifiers and negation, and certain scope orderings (Evang and Bos, 2013). Moreover, the syntactic environments annotated are very limited, limiting the corpus's potential to train a model of grammatical function as an influence on scoping.

2.1.10 Groningen Meaning Bank

Even if we find Evang and Bos's (2013) method poorly aligned with our purposes, might we not rather contribute a stand-off annotation to the GMB than create a new, unrelated resource? As with OntoNotes, integrating diverse annotations of shared texts is a strength, and we would not have to limit our annotations to a narrow class of PPs or a small set of quantifiers. Text coherence, linear order, grammatical function, lexical realization, and presumed world knowledge would all be represented.

The problem again is genre and density. Evang and Bos state that their method covers 'some of the most common configurations giving rise to scope ambiguities involving universal quantifiers,' but at 456 examples in the meaning bank's million words, this 'most common' is not very frequent. Of the GMB's main genres, the legal documents and descriptions of countries favor generalizations, but the fables and news articles disfavor them. Upon developing a good model of human quantifier scope prediction, we would be glad to use it to contribute a layer to a resource like GMB or OntoNotes, but for use in development we prefer something more dense in quantificational generalizations.

2.1.11 Building data de novo

Finding no suitable dataset extant, we have developed our own. We take excerpts of a few sentences from articles in the Simple English Wikipedia, parse them, and annotate with semantic and pragmatic judgments sufficient to render (what we believe to be) the intended reading in logical form.

For the scope prediction project here described, the main point of interest is to disambiguate the outscoping relationships that are required in order to correctly formalize the truth conditions of the intended meaning. Related secondary concerns include disambiguating implied quantificational force, as well as coreference.[5] But this work does not take place in a vacuum.

Another consumer of the annotated data that affects many aspects of the work is a project to prepare a corpus of formalized generalizations—lambda-calculus representations of the Simple English Wikipedia excerpts. These might themselves be used for background knowledge in a computational model of the world (or some domain of it), or the relationship between them and the natural-language source data might help to train models for correctly formalizing verbal explanations from

[5]In the broad sense: Multiple mentions that refer to the same entity or entities. The terminology is vexed here. 'Coreference' should not be assumed to exclude multiple uses of a variable under the same binding. 'Anaphora' should not be assumed to exclude cataphora, let alone taken as distinct from pronouns as in Government and Binding theory. And although some traditions make an exact distinction between anaphora and coreference (e.g. Kempson, 1977), I write in the computational linguistics tradition, which historically has not (but see Sukthanker et al. 2020 for a recent attempt to bring some clarity).

experts in other domains. In any case, the lambda expression project has requirements of its own in addition to a shared interest in quantifier scope. And both of these projects have emerged from a broader research program to model incremental processing from language to semantics within the human mind. Its influence will be visible from time to time also.

2.2 New corpus overview

This section summarizes the current state of the Simple English Wikipedia scope-annotated corpus.

Simple English Wikipedia is a smaller sister project of the well-known collaborative encyclopedia. Its genre is the same: informative text on many subjects. This offers many generalizations and a broad domain. However, its target audience is non-native users of English, a category to which, with some imagination, we can assign natural language processing computers.

The 'simple English' register is not a controlled language; there is no closed list of words, word senses, or grammatical constructions to which writers are confined, although several such lists are published on the wiki for reference. Editorial guidelines encourage using simple tenses, active voice, and short sentences of a single clause, but the guidelines are not meant to override common sense, and the final definition of 'simple' remains subjective. That is, this is still natural language, and not merely a semantic formalism with an unusually English-like syntax.

2.2.1 Two generations of documents

Each document in the corpus is the beginning of a Simple English Wikipedia article. The corpus includes two 'generations' of documents.

First-generation documents are sourced from among the earliest-created articles. Articles that had not grown past two sentences were skipped. Articles of three or four sentences were taken in their entirety. Longer articles were truncated to three sentences (or rather, three segments, counting any standalone noun phrase used as a header). I annotated scope and coreference in these documents, constituting just over 1000 sentences.

	Tokens	Segments	Sentences	Documents
Superseded 1st-gen	2876	180	180	59
Remaining 1st-gen	14,708	844	843	279
Current 2nd-gen	22,866	1435	1373	256
Current	37,574	2279	2216	535

Table 2.2: Summary of fully annotated corpus

Second-generation documents retain the three-segment minimum, but run as long as six segments where possible. The annotation guidelines have also been expanded, reflecting lessons learned from experience and from an inter-annotator agreement study (described in Chapter 3).

Second-generation source articles are selected by the frequency of their (single-word) titles within the text of the Simple English Wikipedia, in descending order. When this criterion overlaps with the earliest-creation criterion, the first-generation document is superseded in favor of the second generation's longer excerpts and better-informed annotation guide.

2.2.2 Size

Table 2.2 describes the fully-annotated data currently available. The count of tokens includes titles and headers; the count of segments excludes titles; the count of sentences excludes both.

This represents the documents that have been annotated for coreference, scope, and implied quantificational force, and have had their parse tree hand-checked and corrected. However, 24 of these documents (100 segments) could not pass the data preparation pipeline, either at the syntax-to-semantics stage or during subsequent semantic processing (for details, see Section 4.2.1.3), and had to be dropped, leaving 511 processable documents (2179 segments) to be divided among training, test, and validation sets (Section 4.2.3).

2.2.3 Pending expansions

The corpus is to be expanded with additional materials now being annotated. A large supplement is now in progress, but was not ready for use at the time the prediction project's dataset was finalized.

This supplement totals 22,924 tokens, 1464 segments, and 265 documents, which will roughly double the available second-generation content. It supersedes 34 first-generation documents.[6] Scope and coreference annotations are done and have been manually reviewed. Hand-correction of syntax trees and annotation for quantificational force (e.g. generic versus existential) are still underway. These are required both for an automated final check that the annotated scopes and coreferences come out acyclic, and for this project's data preparation pipeline, so these documents are presently unreleased and unused.

Coreference annotation has now begun on a second supplement of roughly the same size. Scope annotation is pending further training of new workers.

2.2.4 Density of scopal interaction

Table 2.3 shows the distribution of segments by count of scopally interacting quantifications, within the processable documents.

The 'any quantification' column groups segments by the number of scopally interacting quantifications they contain, and measures the groups' relative size. Those in the 'one quantification' group must interact with a quantifier in another segment. For sentence segments, an interaction is any scopal relationship that, if reversed, would produce a proposition with different truth conditions; for noun phrase segments, it is any scopal relationship that, in some possible world, produces a different restrictor set than its reverse. Many of the interacting quantifications counted are quantifications over eventualities, such as when the phrase 'two blue guitars' implies that there exist, for each entity in its restrictor set, periods of being a guitar and being blue in which the entity participates.

[6] Compare this with the current corpus's 59 supersessions. There is a trend toward more supersessions with more-frequent words (verified with smaller segments of the ranked list). We have construed both the early Simple English Wikipedians' choices of what to write about and the more recent writers' vocabulary use as collective decisions about what ideas are important and accessible, so the correlation between them, though modest, pleases me.

Note that these claims of existence interact with the outscoping ‘two’, because reversing the scoping implies the very different scenario in which two objects are both the guitar of the same single being-a-guitar. If the phrase were just ‘a blue guitar’, however, they would not have that interaction, and if they had no other interaction, they would not count toward their segment’s total and its position in this column.

In addition to eventualities, this column counts quantifications over possible worlds (such as the epistemic ‘must’) and other similar exotica. Note also the eight segments that each contain a single scopally interacting quantification. This column does not require the two interacting quantifications to come from the same segment.

The ‘entity’ column counts only quantifications that are signified by noun phrases, which limits it to quantifications over entities. Guitars are still eligible to be counted, but being-a-guitar is not. The only other requirement, though, is that the quantification has a scopal interaction; it may be with a quantification of some other kind, which is the case in 420 or more segments that have now moved to the ‘one interactor’ bin, and it may still cross segment boundaries.

The ‘entities’ column adds the further limitation of only counting entity quantifications that interact with entity quantifications, and the ‘same-sentence entities’ limits it to the quantifications whose interactions are predicted in this project. The last of these is the most comparable to the distribution table in Rasmussen and Schuler (2020). Note, though, that none of these columns is limited to counting overt quantificational determiners such as numbers. Bare plurals and even pronouns are still included.

2.3 Document preparation

Documents are excerpted from a 2014 dump of Simple English Wikipedia, which contains over 100,000 articles.

The syntax and vocabulary of the Simple English Wikipedia are limited, by the standards of prose for experienced L1 readers. However, the language is as varied as in any existing scope corpus,

	Any quantification	Entity	Entities	Same-sentence entities
0 interactors	43.6%	40.2%	83.8%	84.8%
1 interactor	0.37%	19.6%	0.78%	N/A
2 interactors	14.6%	12.3%	9.27%	9.18%
3 interactors	9.87%	4.22%	3.07%	2.98%
4 interactors	8.63%	1.93%	1.47%	1.47%
5 interactors	7.57%	1.24%	1.01%	0.96%
6 interactors	4.59%	0.41%	0.46%	0.46%
7 interactors	3.40%	0.18%	0.09%	0.09%
8 interactors	2.71%	0.28%	0.05%	0.05%
$\geq$ 9 interactors	4.64%	0%	0%	0%

Table 2.3: Percent of segments with scopally interacting quantifiers

and more varied than in most, since it was not filtered for particular quantifiers or constructions.

It presents sentences in a discourse context and with information structure encoded in syntactic choices, not just a collection of out-of-the-blue utterances or abstract predications. And it must balance the need for brevity against the need for precision, often by presuming on the reader's knowledge of the world.

For all of these reasons, it is much more representative of natural explanatory language than any of the previous resources. But most importantly, despite its linguistic simplicity it is still a rich source of quantifiers of all kinds.

2.3.1 Text selection

First-generation standards for excerpting documents reflected two concerns. The first was text lengths; many articles in this encyclopedia were very short.[7]

The other was domain breadth. The state-of-the-art QuanText corpus (Manshadi, 2014) concerned a single, small subject matter (text editing), and Schuler and Wheeler (2014) had found that the mere identities of the nouns heading two NP chunks were a powerful predictor of their scopes. It seemed that a few simple facts about meronymic (part/whole) relationships in that domain—files

[7]Many still are. A random sample of 40 suggests that at present, the median length is only about 4 sentences.

consist of lines, which consist of tokens, which consist of characters—added up to a devastatingly effective heuristic.

Our decisions to take only three-sentence excerpts and to use the earliest-created articles had both concerns in view, making many short articles usable but preventing long ones from skewing the subject matter to any one domain. The second-generation selection standards reflect lessons learned from this experience.

2.3.1.1 First generation standards

Requiring three sentences excluded numerous stub articles about small municipalities (most of them created automatically from public records), but it retained important, general topics that simply hadn't had much editorial attention. This gave a large pool of articles to choose from, and so a large pool of domains.

The earliest-created articles would have had the most opportunity to be expanded beyond a sentence or two. Moreover, although articles are created by volunteers following their own interests, we expected that many of the people who would volunteer to write a new encyclopedia would take an interest in covering general knowledge broadly.

The wiki's page creation history provided a list of 459 articles that had existed before Simple English Wikipedia migrated to its current software platform, which did appear sufficiently encyclopedic to confirm our expectation,[8] and some 423 of these met the three-sentence minimum (see Table 2.4 for a sample).

As a final check before adopting this list of articles as our source pool, I verified that few of their domains are as meronymic as the QuanText domain is. From their entire content of 13,444 sentences, I extracted every word for which WordNet (Fellbaum, 1998) had one or more noun senses, and queried WordNet for a part/whole relationship in each combination of senses of each pair of words appearing together in one sentence. This is a liberal estimate of meronymy, since

[8]The list is not totally devoid of hobby horses. Early contributors seem to have taken particular interests in Hawai'i, Web technology, and varieties of sausage.

Astronomy	Infinity	Pet
Archaeology	Knowledge	Paradox
Biology	Leap Year	Right angle
Brazil	Montreal	Ranch
Classical element	MediaWiki	Spache Readability Formula
Cooking	Mustache	Sport
Capitalization	Metaphor	Salami
Earth	Mercury (planet)	Soul
Fine	Molecule	Temple
Good	Fishing net	Vocabulary
Green	North America	Microsoft Windows

Table 2.4: Sample of first-generation article titles

there is no guarantee that either word is used as a noun at all, much less in the particular sense needed. But even by this generous measure, only 698 sentences, or about one in nineteen, contained a possible part/whole pair.

2.3.1.2 Second generation

Working with the first generation of documents revealed that our source articles have fairly stereotyped beginnings. The first sentence usually begins with the article's title and consists of a definition. The second and third sentences are often dedicated to explaining parts of it. Longer excerpts in the second generation are intended to capture a wider variety of discourse relations and information structures than before, as well as more facts about the article's subject.

The question of article selection returned also, because clearing the 1000-sentence mark had consumed 338 of the original pool of 423 articles.[9] We were going to need a new way of locating well-developed articles on a wide variety of subjects.

Since the inception of Simple English Wikipedia, the Wikipedia community has developed consensus lists of essential articles, which can be used to guide an encyclopedia's growth or check that its coverage is appropriate to its size. We experimented with their list of 1000. However, we quickly

[9] In fact, the other 85 were also excerpted to three sentences and annotated for scope and coreference, but were abandoned without syntax review and other annotations when the second-generation standard turned out to be much more satisfactory.

Building	Energy	Love
Mother	Lake	Village
Art	Professional	Pakistan
Storm	Computer	Canada
Republic	Star	Mean
Role	Force	Ice
God	Something	Championship
Hockey	Minister	Empire
Battle	Summer	Space
Military	China	Wife
Information	Want	Cup

Table 2.5: Sample of second-generation article titles

found that the list was skewed toward proper nouns, in the form of historical figures, locations, artworks, and so forth, at the expense of generalizations.

After this false start, we based the selection on word frequency. A sample of titles is in Table 2.5. The sample is visibly skewed by the wiki's numerous articles about geographic features and sports celebrities; television, film, and popular music are other topics where it has heavy coverage. Homographs occasionally intrude; the frequency of 'something' is doubtless because of the indefinite pronoun, but the article is about the song by George Harrison, and the article 'mean' is about averaging, not denoting. Nevertheless, widely used words have also brought a broad sample of general-knowledge topics to the fore, and proper nouns are largely avoided.

2.3.2 Syntactic preparation and annotation

Though not of primary concern for this project, the corpus's syntactic annotation determines which data we can prepare and will be described briefly.

The article excerpts are automatically segmented and parsed, using the Petrov and Klein (2007) parser trained on the Nguyen et al. (2012) reannotation of the Penn Treebank (Marcus et al., 1993), into a generalized categorial grammar markup. Scope/coreference annotators then designate outscopers and antecedents, while the parse tree is reviewed separately.

The first concern of that review is the tree's structure. Modifier attachment errors are relatively common and require some rebracketing. Segmentation and even tokenization errors occur from time to time as well; these are coordinated with the scope/coreference team so that the two annotations can be merged correctly when both are done. The larger concern, though, is at the syntax/semantics interface.

The generalized categorial grammar markup distinguishes composition operations for arguments, modifiers, and various non-local constructions such as filler-gap constructions, and each composition operation defines constraints to be placed on a restrictor or nuclear scope set. In general, meanings of modifier predicates constrain restrictors of quantifiers associated with modificands, and meanings of non-modifier predicates constrain nuclear scopes of quantifiers associated with arguments. These marked-up operations are then used to define a set of elementary predications (Copestake et al., 2005) over variables in restriction and nuclear scope expressions for each quantification over entities (typically expressed by a noun phrase) or eventualities (expressed by some verbs, adjectives, and prepositional phrases).

The syntax review therefore includes passing this markup through an automated implementation of the above definitions to ensure that valid elementary predications can be obtained, and editing as necessary.

Loosely speaking, then, one annotation job marks up the pragmatic end of semantics, and the other marks up the syntactic end. There are certain overlaps and exceptions. The coreference task as defined for annotators includes marking the antecedents of reflexive pronouns, although that could be deduced from their syntax, and pragmatic judgments about implied quantificational force fell to the syntactic review to avoid a disruptive retraining of scope annotators.

When a document has been annotated both ways, the pragmatic annotation tags can be automatically transferred to the parse, and the elementary predications can be assembled into lambda terms with the algorithm of Schuler and Wheeler (2014).

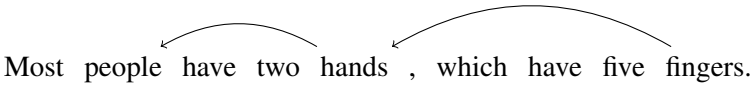

Figure 2.1: Sentence with scope arcs

2.4 Pragmatic annotation task

The annotation process is designed to safeguard accuracy while reducing mental labor.

2.4.1 Scope as dependency

Discussions of scope are often conducted in terms of fairly complex formalisms: Predicate logic following Montague (1973), and/or covert syntactic movement followingMay (1977). Our annotation instead uses a lightweight representation of dependencies (labeled directed graph edges, also called 'arcs') within the original natural-language text. Annotators can visualize, reason about, and discuss scoping judgments in a form analogous to Figure 2.1, then mark them up on preterminal nodes of the parse, reusing skills they learned for annotating coreference chains.[10]

Intuitively, coreference dependencies are pointers from a noun or pronoun to (the head word of) the last prior mention of the same entities/entity. Outscoping dependencies, pointing from one (low) noun to another (high) noun, specify that a set of entities described by the low noun relates to each entity described by the high noun. Behind the scenes, we maintain a convention that maps heads of noun phrases, modal auxiliaries, and negations to quantifications and entities in logical form, so that these word-to-word dependencies can be used as the skeleton for building a generalized quantifier expression like Figure 2.2, but from the annotator side the task largely amounts to drawing arrows.

[10]The sparsity of this representation may seem more reasonable after considering the overlap between any underspecified scope representation and one of its fully-scoped extensions, e.g. the Hole Semantics formula for 'Every boxer loves a woman' diagrammed in Blackburn and Bos (2005, p. 134) and the direct-scoped logical form on the facing page. All of the predicates, logical operators, entity variables, and binding quantifications are duplicated; the difference between them, the 'plugging', is very small.

$$\begin{aligned}
&\textsc{most}\ (\lambda_x\ \text{person}\ x)\\
&\quad(\lambda_x\ \textsc{two}\ (\lambda_y\ \text{hand}\ y)\\
&\qquad(\lambda_y\ \text{have}\ x\ y,\\
&\qquad\quad\textsc{five}\ (\lambda_z\ \text{finger}\ z)\\
&\qquad\qquad(\lambda_z\ \text{have}\ y\ z)))
\end{aligned}$$

Figure 2.2: Lambda expression with the same scoping; generalized quantifiers in small capitals

After marking the dependencies, annotators discuss any concerns in a review meeting and run their work through automated validation, in either order. The validation program checks that neither kind of dependency forms a cycle, and that the annotated scoping suffices for the Schuler and Wheeler (2014) algorithm to assemble a well-formed expression, in which all variable arguments to elementary predications are bound by the lambda abstraction of an enclosing restrictor or nuclear scope set. Any validation failures are hand-corrected and retried.

Formulating the task in this way removes (most of) annotators' need to work with formalized predicate-argument semantics, leaving them free to concentrate on inferred coreference and scope. Keeping the natural-language source text at the center preserves all of its influence on these pragmatic judgments. Doubts are easier to communicate orally, errors are less likely, and error spotting is easier, when other considerations about the correct semantics neither intrude on the question 'How are these quantifiers scoped?' nor are intertwined with its answer. And the scope-as-dependency approach simplifies recruiting and training by reducing prerequisites and allowing us to teach the file formats and computer tools of the job through the coreference task, before tackling the concepts of scope.

2.4.2 The annotator at work

Figure 2.3 shows the beginning of a document, as it might appear on an annotator's computer while at work. At right is the vertical text, with blank lines between segments. The left margin gives each token an address for dependencies to point to. In between sits the syntactic parse, in the

```
0001:(N-lS (N-aD-xN%y|N%y                                                City))

0101:(S-lS (S (V (N-lA (N-b{N-aD}                                        A
0102:    ) (N-aD-lU (N-aD-xN%y|N%y-n0001-yQ                              city
0103:   ))) (V-aN (V-aN-b{A-aN}-xV%is|B%be-xCOPU                         is
0104:    ) (A-aN-lU (N-lZ (N (N-b{N-aD}-x%|                              a
0105:       ) (N-aD-lU (N-aD-xN%|N%-n????-s????                          place
0106:      ))) (C-rN-lR (V-rN (R-aN-rN-lG (R-aN-rN-xR%|A%                where
0107:        )) (V-g{R-aN} (N-lA (N-aD (A-aN-x-lM (A-aN-x                 many
0108:           )) (N-aD-xN%ople|N%rson-yQ-s02                           people
0109:         ))) (V-aN-g{R-aN} (V-aN-lE (V-aN (V-aN-xV%e|B%e             live
0110:             ) (R-aN-lM (R-aN-xR%|A%                                together
0111: ))))))))))))) (.-lM (.-x%|                                         .)))

0201:(S-lS (S (V (N-lA (N-b{N-aD}                                        A
0202:    ) (N-aD-lU (N-aD-xN%y|N%y-n0102-yQ                              city
0203:   ))) (V-aN (V-aN-bN-xV%has|B%have                                 has
0204:    ) (N-lA (N-lC (N-aD (A-aN-x-lM (A-aN-x                          many
0205:       )) (N-aD-xN%s|N%-n????-s????                                 buildings
0206:     ))) (N-cN (X-cX-dX-x%|-yQ                                      and
0207:      ) (N-lC (N-aD-xN%s|N%-n????-s????                             streets
0208: ))))))) (.-lM (.-x%|                                               .)))
```

Figure 2.3: Scope/coreference file, mid-annotation

Generalized Categorial Grammar of Nguyen et al. (2012). It is consulted occasionally, when there is serious doubt whether the annotator's reading of a syntactic ambiguity agrees with the parse,[11] but otherwise plays very little part in this task, so it has been compacted to one line per token, given the smallest meaningful indentation, and spaced away from the text. Its right edge is annotated with tags, each beginning with a hyphen and a lowercase letter and running until the next hyphen or space.

The figure represents a file that has been updated from the automatic parse to the hand-corrected syntax, and so includes syntax-to-semantics annotations beginning with `-x`, which do not concern us here.[12]

Semantic placeholder tags of the form `-n????-s????` are automatically attached to every nominal preterminal category when the file is formatted for annotation. Here the annotator has replaced three placeholders with real annotations, but has not yet double checked and deleted the others. Placeholders identify potential sites of coreference and quantification. They are slightly overdeployed, which is why three remain here. 'Place' only represents a predicate applied to the nuclear scope set of 'city', not an entity reference of its own, so its placeholder is a false positive, and the placeholders in 'buildings and streets' have gone unused because they can be annotated jointly on the conjunction. However, the placeholders have greatly reduced the problem of overlooking scopables, which was noteworthy in the first-generation documents (see Section 3.4.1.1).

The `-n` half of the placeholder represents inheritance, a mechanism for formalizing discourse anaphora and the maintenance of knowledge. It is used at `0202` 'city', pointing back to `0102`, whereby all the facts previously predicated of cities (being a place, being many people's site of living) are included in the new reference.[13] Annotators in fact use `-n` not only for discourse anaphora,

[11] The pragmatic and syntactic annota*tions* must agree, but there is no fixed rule giving any annota*tor* priority. Consensus is usually quick or immediate.

[12] For the curious, lemmatizing tags, with percent signs and vertical bars, set a predicate's name and use. The `-xCOPU` lexical rule ensures that 'a place' will generate a predicate but not a new entity referent.

[13] The title, `0001` 'city', offers no facts for `0102` to inherit. We annotate anyway because there is nevertheless a reader judgment of coreference, and because the title's special status in the document may make the structure of the dependencies itself informative. It is reasonably possible, for example, that the title as a device of information structure affects how later, coreferring noun phrases are read; and it is certainly the case in our documents that the title is strongly associated

but for any non-first mention of an entity or entities. Post-processing re-tags those that are actually second uses of one bound variable, rather than new bindings.

2.4.2.1 Marking scopal interactions

After coreference, annotators perform a pass adding `-yQ` tags. This is a second, more careful form of placeholding to refine the set of words that may need scope annotations. Nominals signifying quantification over entities are our main focus to begin with, but other words such as modal auxiliaries also signify quantifications and can have scopal interactions.

Moreover, not every word that signifies a quantification will need annotation. Neo-Davidsonian semantics (Parsons, 1990) quantifies over eventualities for each elementary predication,[14] and almost all of these quantifiers are existential and low-scoped. Most existential quantifications over entities in our data are low-scoped as well. We have introduced a convention that allows low-scoped existentials to be left unannotated (discussed further in Section 4.2.1.1.1), and the `-yQ` tagging is also in part a process of identifying where the convention applies and dismissing these sites from further attention.

We have found that delaying the step of writing the dependencies by inserting this methodical tagging step produces more accurate work than diving right in. It clarifies the questions at hand by defining where the annotator should and should not look for scopal interactions.[15] It tends to produce a mental sketch of the answers also, so that there is less error-prone backtracking to edit a half-annotated text.

---

with a distinctive pattern of discourse that both uses and mentions the title word (See Section 2.4.3.4.6). Dependencies to the title capture this information.

[14] Akin to representing 'Sandy introduces Kim to Chris' not as introduce(Sandy, Kim, Chris) but as

$$\exists e : \text{introduction}(e) \wedge \text{agent}(e, \text{Sandy}) \wedge \text{theme}(e, \text{Kim}) \wedge \text{patient}(e, \text{Chris})$$

.

[15] Among other things, it relieves a liability of leaving low existentials unannotated, which is that existential and generic assertions compete for the same morphology: bare plurals, mass nouns, and even some singular count nouns with *a/an*. `-yQ` tagging defines a time to consider this when there are fewer distractions. Leaving it to be done in the heat of battle, as it were, leads to generics being overlooked.

With `-yQ` tags flagging where annotatable scope dependencies may begin and end, the last step is to record where they actually do. The `-s` tag appears at the lower scope and contains the address of the higher. The figure shows this step half-done, with a tag on `0108` 'people' but none yet on `0206` 'buildings and streets'. The `-s` tag shown abbreviates the address to two digits for a token within the same segment, a provision also available for `-n`; some annotators use it, while others prefer four-digit addresses everywhere. For historical reasons, it is not customary to remove `-yQ` when an `-s` tag is placed, but it is not used in further processing of the data.

2.4.3 The annotator instructed

This section describes some distinctive points of the training materials for pragmatic (and especially scopal) annotation.

2.4.3.1 Intended reading

Annotators are asked to mark the coreference and scopal interactions of a document's most probable intended reading. Training materials illustrate some of the cues that may make a certain reading more natural in its context. For example, in

(2) All lizards have scales. Gila monsters' scales are pink and black.

we are probably meant to understand that Gila monsters are a kind of lizard. Elements that may play into this judgment include

- Information structure: 'Gila monsters' is leftmost in its sentence, a position often given to things that are familiar.
- Metadiscourse: A connective like 'specifically' or 'however' would have shaped what we inferred.
- Presumed relevance: We expect the sentences to share an informative purpose.

```
assume answer "false"
for each foot f:
    if f smells fine, make answer "true"
    if not, don't change answer
stop and report answer
```

Figure 2.4: Robot program verifying an existential quantifier

- Presupposition: 'Gila monsters' scales' presumes that they have them, and we expect that the text justifies this.
- Lexical meaning: 'Monster' connotes fright, which fits many people's feelings about lizards.
- Factual knowledge: The writer may have expected (justifiably or not) that readers would already be familiar with Gila monsters.[16]

For the writer of the text, it might have been wise to make the intent more clear by deploying such cues differently, but for annotators, the example serves to emphasize that they are fair game. *Technically* it doesn't *say* that Gila monsters are lizards, but almost everything we want annotated is something the text technically doesn't say—a conclusion we're not forced to draw, a judgment. So if it looks like we're supposed to read 'Gila monsters are lizards' into the text, then annotators should write it in (with an inheritance dependency).

2.4.3.2 Teaching scope: quantifiers as robot programs

The basic practice, traceable to Tarski (1933), of interpreting quantifiers by iterating over individuals in the universe can be extracted from symbolic logic and deployed to create an intuition about scope. Annotator training materials use the analogy of programming a very durable robot to verify, for example, 'Some foot smells fine' by examining every foot in the world (see pseudocode in Figure 2.4).

[16] In an online text, the expectation could be justified by making the phrase 'Gila monster' a link. Because our corpus strips out most wiki markup, it is blind to hypertextual rhetoric like this, but we haven't missed it much. Only about one sentence in a thousand has been so questionably relevant as to drive us back to check the source article, and more of these are explained by seeing sentences that were truncated than by seeing hyperlinks.

```
assume every-kid-answer "true"
for each kid k:
    assume this-kid-answer "false"
    for each table t:
        if k is standing on t, make this-kid-answer "true"
        if not, don't change this-kid-answer
    if this-kid-answer is "true", don't change every-kid-answer
    if not, make every-kid-answer "false"
stop and report every-kid-answer
```

Figure 2.5: Robot program verifying an existential quantifier

This may amount to replacing one formalism with another, but the idea of an algorithm is culturally current in a way that symbolic logic is not. The training materials next demonstrate a universal quantifier in an analogous format, as a program for verifying 'Every Gila monster is venomous' by assuming truth and searching for a counterexample.

A program is then proffered for verifying 'Every kid is standing on a table' (Figure 2.5), where a kid is identified as a counterexample not by checking a single simple property, but by running an inner loop that examines tables as they relate to that kid. The exact wording of the previous pseudocode is designed for this scenario, to keep both quantifier loops recognizable and their nesting clear.

This program corresponds to the intuitively preferred direct scoping of the sentence. Once it is understood, a program for the inverse scoping is offered as an alternative. The discussion shows how it is also a reasonable approach to the sentence, then leads around to highlight some possible worlds where they don't produce the same answer. Since the two programs are extremely similar, the difference between them is easy to reduce to which set of entities gets searched by the outer loop or the inner, and this is scope itself.

2.4.3.3 Teaching scope: bag of tricks

For annotators better acquainted with computer programming, a final robot program cements the relationship between how linguists talk about scope and how programmers do: `if g is biting k,`

`report "Danger, Will Robinson!  Danger!"` is well-intentioned, but the robot should reply 'That does not compute'. The variable names are not defined to refer to any particular entities, when no `for each` statement is walking them across the universe. They are meaningful only within the loop bodies, and 'scope' to a programmer is the portion of a program where a name is meaningful. So the programmer scopes exhibit the same nesting as the linguist scopes.

A series of useful heuristics is provided also. Paraphrasing an assertion to begin 'for each X, for that X's set of Ys' is often easy to do and brings the intended scope forward. Attempting to substitute determiners may help: 'each' seeks wide scope, 'the same' avoids narrow scope, 'one or another' tends moderately toward narrow scope. Follow-up sentences with singular or plural pronouns may be compatible only with one scoping or another.

A challenging but powerful last resort is to devise possible worlds where the scopings have different truth values, and ask which is the kind of world the writer wants readers to believe in. Between the natural credibility of each world and their fit to the writer's point, there is usually a distinct difference.

2.4.3.4 Selected other guidelines

Annotator reference materials covered a range of relevant fine points and challenges. The most important of these was not to annotate a dependency where there is no scopal interaction, i.e. for an outscoping that can be reversed without changing the conditions that make the statement true or false.

2.4.3.4.1 Indifferent scope Consider one universal quantifier immediately outscoping another:

(3) Every Hatfield hated all McCoys.

As a matter of deductive logic, either quantifier can be given the wider scope. Both formulations will be true, or both false, under any given set of facts, so the quantifiers' relative scope is a matter of indifference; they cannot interact scopally. The same is true for two adjacent existential quantifiers:

(4) Some singer shot some sheriff.

Though this case is vastly more common, it mostly occurs among existential quantifiers at the narrowest scopes, where the previously mentioned convention (Section 2.4.2.1) suppresses annotations anyway. On the rare occasions when non-interacting quantifiers appear at wider scopes, special markup is needed to accommodate corpus users' different needs (described in Section 4.2.1.1.2), but just omitting any annotation is sufficient to make validation fail, ensuring that the document will come up in a review meeting.

2.4.3.4.2 Noun/referent mismatches Another special circumstance is nouns that do not introduce an additional referent. Complements of copulas, appositives, and certain quasi-appositive postmodifiers (such as in 'the Republic of Austria') signify predications, which do not (typically) participate in coreference chains or scopal interaction, so annotators are warned to pass over them and use their subjects/heads instead. Phrases like 'a lot of' typically serve as multiword quantifiers, in that there is no 'lot' available for anaphora afterward (real estate and auctions being the principal exceptions).

'For example' and similar metadiscourse are an adjacent case; they do refer to entities, but the entities exist with us, the readers, not among the things the rest of the text describes, and so very seldom have any referential or scopal interactions.

On the other hand, annotators occasionally work with entity referents other than directly through nouns. Annotations on a conjunction represent duplicating the annotation on each of the conjuncts, which downstream software can spell out in full.

When nouns are elided, guidelines recommend a substantive adjective or determiner as proxy. Our software does not include general-purpose syntax-to-semantics rules for elision, but the validation failure brings the document in for review, where the proxy annotation reports the coreference and scope judgments, so that the correct logical form can be patched in by hand.

2.4.3.4.3 Anaphora to restrictor sets Lastly (for the present discussion), guidelines provide for a few species of anaphora beyond simple coreference, with two special notations.

(5) Most feet stink, but some of them smell fine.

In this sentence, contrary to ordinary coreference, 'some of them' should acquire only footness from its antecedent, not stinking also. In logical form, this is done by referring back to the restrictor set of 'most' rather than the nuclear scope set as usual. Outscoper dependencies may also need to target the restrictor, as in

(6) Most bears with three paws are exceptionally dangerous.

If the elementary predication of having paws uses the bear variable bound by the nuclear scope set, the generalization is wrong: The logical form would then turn out to say that most bears have three paws and are exceptionally dangerous. To restrict the generalization to three-pawed bears, that predication must use the bear variable bound by the restrictor set, and so the paw variable it uses must be within that scope. The notation for a dependency to a restrictor set is the letter `r` suffixed to the address.

2.4.3.4.4 Bridging anaphora and presupposition The other special notation is a catchall dependency type for pragmatic judgments about justified presupposition, several of which have the effect of making a referent available.

Bridging anaphora, as seen in Example (7), might be analyzed in terms of a special class of nouns that have an implied argument.

(7) Each bus stopped at the border and the driver got out.

On such an analysis, 'driver' has a patient argument like verbs do, but it is pragmatic judgment rather than syntax that points it toward 'bus'. That analysis is less comfortable for discourses like

Example (8).

(8) Chris's shower stall fell through the floor. The joists were rotten from water damage.

We do not want to propose that *joist* is a noun of a special kind, with an implied argument for the floor it is a part of. Nearly any noun can trigger bridging anaphora, not just a special subclass, and the relationships between these nouns and their antecedents are more like the numerous relationships of noun–noun modification (for example, see the taxonomy of Tratz and Hovy, 2010) than like a small set of syntactic theta roles.

This suggests modeling bridging anaphora in terms of weak familiarity (Martin, 2012): A definite noun phrase like *the joists* projects the presupposition that its referent is uniquely identifiable. We supply a dependency back to a prior referent that justifies the presupposition, if finding that referent is a pragmatic judgment (though not if a syntactic device like *of*-modification makes the connection clear). This is in the spirit of 'presupposition as anaphora' (van der Sandt, 1992).

In Example (7), we would annotate an outscoping parallel to the presupposition dependency, which is a frequent pattern in bridging anaphora,[17] but other presupposition triggers often stand in some other relationship to the justifying referent. For example, words like *another* and *later* presuppose an identifiable prior entity or point in time.

Admittedly, this dependency (which we mark with `-w` for 'weakly familiar' or 'weird anaphor') has barely a shadow of a meaning. It shows that we have read into the text some kind of semantics that both referents participate in, but it is silent as to their structure and truth conditions. To classify its uses more narrowly, describe those semantics, and predict them falls beyond this project's ambit.[18] But the pragmatic phenomena it stands for, inexactly described as they are, routinely touch

[17] Further study of this pattern should also examine it an extreme case of implied domain restriction or of quantifier subordination.

[18] To quote William, 'It's not ideal, but it postpones the complication until somebody else's thesis.' A reasonable step in that direction might be to predict the exact semantics of noun–noun coordinations, which are similarly underspecified at present. The success of neural methods in that area (e.g. Dima and Hinrichs, 2015) and in intricate structured prediction (e.g. Prange et al., 2021) suggests that revisiting `-w` need not be a far-future project.

upon the ones we are presently concerned with, and can even be confused with them. This makes recording it useful anyway.

For annotators, marking the catchall dependency is not just a more rewarding action after discovering a subtle pragmatic relationship than forgetting about it (although it is that). It also unburdens memory of a factor that coreference and scope must stay consistent with, it may even provide some insight about which variables need to be in scope where, and it fills the annotation site with a record of the conclusion that the relationship is not coreference or scope, so that it cannot be mistaken for an omission on later re-examination. Meanwhile, for future researchers, it flags interesting phenomena for followup.

2.4.3.4.5 `-w` for slippage Simple English Wikipedia writers are often willing to presume, when they have introduced a whole, that its parts are identifiable, and vice versa. Having introduced either, they slip to the other and back, trusting the reader to fill in the connection with world knowledge. They do the same with such distinctions as use/mention, group/member, or type/token. The `-w` presupposition dependency helps annotators to document how they resolve this fluid usage into exact semantics.

2.4.3.4.6 Use/mention A non-trivial fraction of our documents has slippage between ordinary *uses* of a word or phrase that refer to some set of other entities and occasional *mentions* that refer to the word or phrase itself. Simple English Wikipedia writers are generally not scrupulous about using quotation marks or other formatting to distinguish the two, which increases the chances that an unprepared annotator will incorrectly mark them as coreferring. Training materials warn against this by illustrating that the word 'potato' cannot be fried golden brown, the vegetables do not consist of six letters, etc., so we have to establish two separate coreference chains to avoid mixing up the facts.

Sometimes wiki writers expressly establish both the relationship and the distinction between

word and object, using verbs like ‘to mean’ and ‘to be called’.[19] In this case, the foregoing advice is enough to get a proper annotation.

Often, though, writers take the relationship for granted, and produce texts like

(9) Potatoes come from South America. They are an edible tuber. The word is borrowed from a Carib language via Spanish.

The definite description ‘the word’ presupposes identifiability, which is justified by the previous use ‘potatoes’. In the general case, this is what the `-w` dependency is for. It stands for the understood eventuality of meaning or being-called, in which the vegetables and the word participate (modulo some adjustments for lemmatization).

We have made special provisions for annotating the most common case of slippage, but if they did not apply, the annotator would have to make sure to point the dependency to the restrictor set of ‘potatoes’, because the only thing that determines whether a vegetable participates in that eventuality is whether that use of the word referred to it. Pointing to (the nuclear scope set of) ‘they’ would pick up all of the later asserted facts, amounting to a claim that we are thinking of the word ‘potato’ not because we read it, but because we read a pronoun that referred to the edible, South American subset of potatoes. The annotator guidelines cover such details.

However, in our documents, nearly all use/mention slippage is limited to the document’s title and its referent. Construing a title like ‘Potato’ (somewhat generously) to refer to the entire set of things that can be called ‘potato’, with no other facts yet predicated to constrain it, the set contains both vegetables and utterances. Both coreference chains can inherit from the title and add a predicate to take a subset containing only potatoes or only ‘potato’s, which keeps them separate and gives this distinctive kind of discourse a distinctive bifurcated inheritance structure. Pointing the

[19] ‘To mean’ is quite reliable: Words mean entities. ‘To be called’, unfortunately, serves dual duty as a fancy copula. In ‘I was called “Sūn Wěixiáng” in Taiwan’, an entity is called words, but ‘[Repeated daily] traveling is called commuting’ doesn’t just express a fact about how people speak. It names a predicate, which is why later reference can just say ‘commuting’ and not ‘what is called “commuting”’.

dependency to the title makes it unnecessary to attend to the careful distinctions discussed in the previous paragraph.

2.4.3.4.7 Group/member, type/token A group, population, collection, or type can be referred to its own right, and things can be said of it that are not true of its members:

(10) a. The Mudville Ballet is large and uncoordinated.

b. All of the dancers are small and graceful.

In this example, terminology makes the distinction fairly easy to maintain; I have conveniently not named the organization the Mudville Dancers, and no one person who dances can be called a ballet. Elsewhere, the terminology can obscure equally valid distinctions:

(11) a. The Beatles were men of flesh and bone.

b. The Beatles dissolved in the 1970s.

(12) a. In the whole language there are only seven words you can't say on TV.

b. In George Carlin's 873-word routine, there were 46 words you can't say on TV.

Making the distinction is not our annotators' primary mission, but sometimes it affects correct coreference, or has to be clear before trying to reason about truth conditions for scoping.

Clouding the issue, ordinary language includes making claims apparently about a group that actually apply distributively to the group's members:

(13) This line of computers had a built-in floppy drive.

We have to read 'this line' to refer to the computers produced in said product line, to make the floppy drives come out right. A product line is an organizational abstraction in which hardware like a floppy drive cannot be mounted, let alone the many drives used in the manufacturing run.

Similarly, phrasing like ‘that kind of X’ is short for ‘Xes of that kind’ when not discussing ontology or type theory. And discourses easily slip across the divide:

(14) a. There are two types of albums, studio albums and live.

b. Sam owns 65 studio albums and only two live.

We have made it a rule to take overt language at face value where possible, so the first of these sentences would be taken to genuinely discuss types, but the second is clearly iterating through the album tokens that inhabit them. Then the first use of ‘studio albums’ signifies the one type that participates in the eventuality being-the-type-inhabited-by-studio-albums, but the second use signifies the many tokens that participate in eventualities of being-a-studio-album.

Just as with use/mention, annotators could use the catchall pragmatic `-w` here to hint that there is a distinctive relationship, but not identity, between the type and token referents, and in particular to record that knowing the type makes the tokens identifiable. But type/token often has an additional complication. Apart from our finagle with article titles, a given word in the documents is almost always unambiguously a use or a mention. Simple English Wikipedia writers freely elide the repetitions and stipulations that would make this true for type/token.

The guidelines for assigning group/set/type or member/element/token status are, in brief:

1. Being exact about type/token may require referents for both, and the type and token referents may both have good claims to be represented by the same word. The word-and-dependency notation cannot handle this case. Make an informal note and bring it to our attention.

2. Align with overt wording such as ‘type, kind, set, sort, group, species’ wherever it can reasonably be done, but depart from it wherever necessary.

3. In the absence of such overt wording, try to construe the semantics as predicating of tokens, and leave it to our ubiquitous quantifications to represent claims about the type. Again, apply this guideline so far as is reasonable, but use predications of types where necessary.

4. When you can unambiguously assign words to type and token status, avoid `-n` inheritance between the type and token. Add `-w` if the relationship between them is unstated.

Examples (15) through (17) further illustrate the slippage, with sentences first as found in the source documents, then rewritten for type/token rigor.

(15) a. [The Nēnē] gets its name from its soft call.

b. The type 'Nēnē' gets its name from its tokens' soft calls.

The type reading is required because under an all-token reading, a Hawai'ian goose cannot be called a nēnē until it calls, and because the species, which was under discussion, really does have the name in its own right.[20]

(16) a. The macadamia nut is the fruit of a tree that first came from the east coast of Australia.

b. A generic macadamia nut is the fruit of some tree of a type whose tokens first came from the east coast of Australia.

The nuts are still macadamias even if grown on a tree that has never been in the Southern Hemisphere. The logical form here will be more than our annotator-friendly word-based front end can handle, with three tree referents: The tokens on which the nuts grow, the type, and the tokens originating in Australia.

(17) a. Hydrogen is the true primordial substance, the first atom produced after the Big Bang.

b. A portion of hydrogen is a portion of the true primordial substance, a set of atoms of the type of the first atom produced after the Big Bang.

'Substance' could be construed as referring to a type, but the semantics for connecting tokens to the type would have to be shoved into the copula. In any case, the mass noun 'hydrogen' must

[20] Means could of course be contrived to propagate the name nēnē to the type when it pertains to enough of the tokens, but that's a lot of tacit mechanism.

be bridged to its countable atoms across the apposition with more finesse than is built into our compositional syntax-to-semantics rules.

As these examples show, type/token slippage creates a streamlined, 'simple' phrasing with an intricate relationship to the exact type/token semantics. That intricacy goes beyond what our annotation system can express, which is perhaps unsatisfying, but also unsurprising and (for present purposes) unproblematic.

Type/token proved 'very challenging' for the QuanText team 'in spite of choosing a specific domain with fairly intuitive quantifier scoping' (Manshadi et al., 2011), and predictably did not get any easier in the encyclopedia. However, resolving it in detail was rarely necessary for our direct purpose of marking annotation and scope, and recording it was, like other uses of catchall `-w`, an error reduction device and a piece of opportunistic future-proofing.

No, it could not record everything we figured out in our occasional forays into type/token resolution. In that sense, some effort has been lost. For any hypothetical future effort to exhaustively annotate type/token, what was lost is a drop in the bucket. For present uses of the corpus, only the coreference and scope dependencies are directly relevant, and the impact of `-w` lies in the erroneous coreference and scope annotations it helped to prevent. In that sense, we have the type/token annotation we need.

2.4.4 Summary of pragmatic annotation

The annotator-facing task for both coreference and scoping largely amounts to drawing arrows on the original text, omitting a large number of highly predictable scopings for quicker work. This lowers barriers to entry, spreads out training, and supports using a full range of pragmatic cues to inform coreference and scoping judgments. Training materials teach about scope from this perspective, without including the entire apparatus of formal semantics.

Annotation placeholders, a methodical sequence of tasks, and automatic validation serve to further highlight what needs to be annotated, safeguarding accuracy and completeness. Reference

materials address the most frequent complications in the areas of determining scope, annotating it on the text, and distinguishing scope and coreference from other, related pragmatic judgments.

2.5 Proper nouns

I took part in the other, primarily syntactic annotation on a limited basis, after we found our lambda terms were ascribing too many scopal interactions to proper nouns. As a practicality, we handle them with essentially the same syntax-to-semantics machinery as common nouns, which allows them to scope,[21] but they should not have scopal interactions, if they are constant references to a single entity—not contingent on how an outscoping variable is assigned, nor offering alternatives for an outscoped variable's binding to be contingent on. Therefore their scoping is indifferent.

In practice, single reference is an oversimplification. Names of people, places, and even calendar entities like months and years are not always univocal; we disambiguate by finding a candidate referent that is relevant to the eventualities and co-participants mentioned in context. This is much like bridging anaphora.

I ran through the syntax annotation files and added a tag to take proper nouns down to the narrowest scopes, where the set of co-participants is best filled out, and to prevent the vast majority of their scopal relationships from being identified as truth-conditionally relevant during the data processing pipeline. I tagged 333 proper nouns in first-generation documents, 420 in the second-generation documents used in this study, 382 in the documents now passing through the rest of their syntax annotation, and 254 in documents just now entering coreference annotation.

[21] Although there is a philosophical debate about whether names refer as descriptions of some sort, or as pointers directly to an entity (see e.g. Michaelson and Reimer, 2019), it is fairly remote from our interests in generalizations and quantification. As a practicality, we act as though there is a predicate for being Kermit the Frog just as there is a predicate for being green. From this perspective, the debate amounts to asking whether that predicate is our summary of some more complex intensional expression, or just the indicator function of an entity set.

Chapter 3

Inter-annotator agreement

This chapter reports work done to devise an appropriate measure of inter-annotator agreement, scores of first-generation data on that measure, and an analysis of the reasons for annotator disagreements. This analysis shaped additions to annotation procedure and annotator guidelines adopted for the second generation of documents.

The design of the measure and the analysis of errors include assumptions and semantic modeling decisions that have since been changed. The most prominent of these is miscellaneous existential quantifications floating to the topmost scope, since this is scored in the agreement measure. These are now scoped low. Semantics for pronouns, copulas, appositives, and relative clauses have also been rethought. They are reported here as they were understood during first-generation annotation.

A human-annotated scoping of a document is highly structured, in that we can expect it to be transitive and acyclic, treat multiple mentions of an entity consistently, and so forth. One can compare two such annotations by decomposing them into 'atomic' scopings between pairs of entities, but the global, structural concerns just mentioned mean that the atomic scopings are not independent of one another. Merely counting agreements among atomic scopings risks mis-stating the degree of agreement between two annotations. To address this, I have developed a chance-corrected measure of agreement.

Two annotators produced independent markups of a random sample of 33 documents (99 sentences), none of which they had previously discussed. Their titles were seen previously in Table 2.4.[1] Agreement between the annotators was calculated with the chance-corrected measure, and

[1] Three of these documents have since been superseded by second-generation annotations: Earth, Good, Green.

for the sake of comparison was also calculated with measures reported in Manshadi et al. (2011) for QuanText.

3.1 Measures of Manshadi et al. (2011)

In QuanText, a 100-sentence sample was annotated by three annotators for IAA. Agreement was calculated only over scope relations, not coreference, because they considered the coreference annotations very easy to produce and did not want them to artificially boost the agreement score.

For comparison, each scoping of a sentence was considered as a directed acyclic graph (DAG), its nodes being the scope-bearers (NPs, negations, and other operators) and its arcs being the annotated outscoping relations. Outscoping is transitive; if A outscopes B and B outscopes C, A necessarily outscopes C. Therefore an arc between A and C can be added without contradiction. Adding all such arcs to the graph expands it to its transitive closure.

Now each pair of scope-bearers can be labeled. If an arc exists between them in the transitive closure, they interact scopally, and the label is either *direct* or *inverse* according to their word order. Otherwise, the label is *no interaction*.

Manshadi et al. defined *constraint-level* agreement by comparing the labels of individual pairs in different scopings, and *sentence-level* agreement as agreement on every pair that can be generated from the sentence. For each level, they used Fleiss's κ (Davies and Fleiss, 1982) as a chance-corrected measure of agreement.[2]

They also defined a variant measure, κ-EZ, which considers two labels to match when either of them is *no interaction*, as well as when they are the same, increasing the level of observed agreement. The generalization from constraint-level to sentence-level agreement is all-pairs matching as before. Unsurprisingly, κ-EZ scores were much higher than plain κ.

[2]The measure is misreported as 'Cohen's κ for multiple annotators', but as pointed out by Artstein and Poesio (2005), Fleiss's κ actually generalizes not Cohen's κ but Scott's π (Scott, 1955), .

3.1.1 Criticism

The notion of comparing transitively closed DAGs by comparing labeled pairs of nodes is an important one to which we will return, but as applied here it is problematic. Fleiss' κ is unsuited for the constraint level, generalizing to the sentence level by total agreement loses information, and the κ-EZ variations overstate reliability. We take each point in turn.

To be meaningful, chance correction requires an accurate model of what can happen by chance. κ's chance model assumes that every item compared is independent of every other, but this is not true if our items are single arcs.

The presence and direction of each outscoping arc are strongly constrained by other arcs in the graph. The arcs that are added in the transitive closure are determined by the elementary arcs in the raw annotation. The annotator observes local and global requirements: Anaphors and cataphors must fall within the scope where their antecedent is bound; the graph as a whole must be acyclic.

Violating κ's independence assumptions in this way would be particularly a problem for us, because the larger the graph, the more such constraints there are. Moreover, we expect users of a scope-annotated corpus to be concerned with the semantics of entire clauses, sentences, and discourses, and not solely with the relative scopes of individual referents. For both of these reasons, the proper granularity for measuring chance-corrected IAA is whole DAGs, not individual arcs.

At the sentence level, a single disagreement overrides arbitrarily many agreements, producing a substantial undercount. Again, this is a larger problem for us since our items are not single sentences (typically 10–20 words in QuanText) but documents (average length slightly over 30 words).

Manshadi et al. (2011) are correct that *no interaction* pairs do not affect a sentence's truth conditions. However, I do not take this to justify counting them as matches to *upward* and *downward* pairs, as in the κ-EZ measure. If the *no interaction* annotator is correct, the other party has asserted a scoping relationship with no truth-conditional effect. If the other annotator is correct, *no interaction* misses a scoping relationship that does affect truth conditions. We would not count either of these as supporting the data's reliability.

	upward	downward	incomparable
upward	24	6	35
downward	6	128	140
incomparable	47	192	10224

Table 3.1: Pair-label confusion matrix.

3.1.2 Comparison

In light of these issues, I determined to find a more appropriate measure of IAA for scope annotations. However, for the sake of comparison with previous work, I have also calculated some Fleiss's κ values. Our confusion matrix at the constraint level is shown in Table 3.1.

Our Fleiss's κ at this level was 0.409, versus 0.750 for the comparable measure in QuanText or a Cohen's (1960) κ of 0.52 in the Higgins and Sadock (2003) data. Whatever methodological concerns one might have had with using this statistic, the outcome strongly suggested expanding our annotation guidelines and performing a second pass through our corpus.

I did not calculate the κ-EZ measure to compare. One reason is methodological, as discussed above. The other is practical: I could not extract sufficient implementation details from the Manshadi et al. (2011) description. They state the effect on observed agreement of counting *no interaction* as a universal match, but not whether or how it is accounted for in expected agreement. It is clear enough that one might merge, say, the *upward/incomparable* cell of the confusion matrix into the *upward/upward* agreement cell, but not clear what happens to the *incomparable/incomparable* cell.

I did calculate a variant measure, which might be considered the opposite of κ-EZ. I knew that *incomparable* results would be frequent, because they arise profusely when the scope DAG branches, as it does in Example (1).

(1) Every car here has four wheels and an engine with four cylinders.

The ordinary scoping is shown in Figure 3.1. *Wheels* and *engine* are both outscoped by *car,* but do

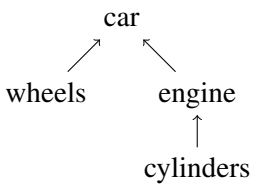

Figure 3.1: Branching scope in Example (1).

not interact with one another. Consequently, any scope-bearers below either of them, as *cylinders* is here, will also not interact, and the number of non-interacting pairs grows quadratically with the depth of the branching.

Fearing that the *incomparable* results might inflate our score, I therefore also calculated agreement over only those pairs that both annotators labeled as interacting. This raised our Fleiss's κ to 0.755, which suggests that our inter-annotator disagreements were most often about the presence of scoping, rather than its direction. On inspection, *incomparable* pairs were generally the result of one of the annotators overlooking a scope-bearer (see Section 3.4). A second annotation pass would benefit from tools to highlight all scope-bearers and so prevent oversights.

In the end, though, all of these variant measures attempt in various ways to work around the fact that scope annotations have internal structure. We turn now to the possibility of addressing that fact.

3.2 Chance-corrected IAA over Structured Items

Artstein and Poesio (2008) argue for computational linguists to measure IAA with chance correction in general, and with Krippendorff's α specifically (Hayes and Krippendorff, 2007). Krippendorff's α defines observed disagreement between two codings of an item in terms of a distance function, and determines expected disagreement by using the same function in an exhaustive permutation test, measuring the distance between codings of *different* items. Crucially, α is agnostic as to which distance function is employed, as long as it is a metric.[3] In fact, by choosing certain metrics one can

[3] A distance function δ is a metric iff it satisfies these axioms for all a, b, c in its domain:

- $\delta(a, b) = 0 \Leftrightarrow a = b$

re-create many other IAA statistics in terms of Krippendorff's α.

Skjærholt (2014) sees in this the solution for IAA over annotations with internal structure—syntax trees, in his case. One finds a distance metric able to embrace the complexity of the annotations, and plugs it into the Krippendorff α formula. The mean squared distance between annotations of the same item gives D_o, observed disagreement; the mean squared distance between *all* annotations, of whatever items, by whatever annotators, gives D_e, expected disagreement, and $\alpha = 1 - D_o/D_e$.

Edit distance measures for trees and other graphs are surveyed in Bille (2005); Zeng et al. (2009); Gao et al. (2010); not all are metrics, however. For purposes of exposition, Skjærholt uses an off-the-shelf tree edit distance metric (Zhang and Shasha, 1989) as his metric, but for real applications he suggests using a distance function adapted to the specific nature of one's data:

'The use of a distance function to describe [α] means that more fine-grained distinctions can be made; for example, if the set of labels on [syntax trees] is highly structured, partial credit can be given for differing annotations that overlap' (Skjærholt, 2014, p. 941)[4]

We have defined a distance function to capture meaningful overlap between scope annotations: We preprocess the annotations to create an explicit scope arc for every scopal interaction. We establish a correspondence between two annotations' scope-bearers, and therefore between their scope arcs. Finally, we use the symmetric difference of the two sets of scope arcs to measure the distance between annotations.

Each step of this process must account for the meaning of the data it handles, in order to credit annotator overlaps properly. The earlier steps are most strongly affected by the nature of the data. The remainder of Section 3.2 will review the considerations at each step and conclude with examples of chance and non-chance agreements in Section 3.2.4.

- $\delta(a, b) > 0 \Leftrightarrow a \neq b$
- $\delta(a, b) = \delta(b, a)$
- $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$

[4]In that vein, SuMoTED (McVicar et al., 2016) might be a better edit distance function for syntax trees. Unlike Zhang/Shasha, it does not insert or delete nodes, and its elementary moves are topologically local, so it traces an intelligible path through the space of possible parses on its way from one annotation to the other.

3.2.1 Preprocessing: Inheritance and Transitivity

Preparing an annotation for IAA testing must involve at least the following steps: Identifying scope-bearers, extracting inheritance dependencies that affect them (some annotated by hand, some supplied by the syntactic parse, and one between each nuclear scope set and its restrictor), inferring transitive inheritance, extracting annotated scope dependencies, locating scope-bearers with no outscoper, propagating scoping through inheritance dependencies (including inheritance paths that run through restrictor sets), and inferring transitive scoping.

Most of these steps are straightforwardly implied by the nature of the data. A few call for some explanation.

The following were identified as scope-bearers: all noun phrases; any term with a scoping arc annotated from or to it; and any term connected by inheritance (in either direction) to a scope-bearer.

Special attention must be paid to propagating scoping through inheritance dependencies, and to scope-bearers with no annotated outscoper.

3.2.1.1 Propagating scoping through inheritance

The bottom of a scoping arc must be propagated down any inheritance chain. The precise meaning of an inheritance arc is that every constraint that applies to its target (e.g. an antecedent) must also apply to its source (e.g. an anaphor). Being outscoped by another referent is one such constraint, so we must infer an outscoping arc from the anaphor to its antecedent's outscoper.

Less obviously, the top of the scope arc must be propagated up the inheritance chain.[5] Example (2) shows why.

(2) I visited each woman there, and she always had a well-fed donkey.

[5]This conclusion, built into the IAA software, rests on the incorrect assumption that the two clauses must use the same variable in the semantics. In our current understanding, 'she' is a discourse anaphor, with its own variable and its own implied quantification, and with the collected facts about 'every woman' defining its restrictor set.

Figure 3.2: Propagating outscoperhood up inheritance chains. Dashed arrows show inheritance, single arrows show annotated scope, and dotted arrows show inferred scope.

The inheritance and scope arcs are shown in Figure 3.2a *donkey* is outscoped by *she,* which inherits from *each woman.* The inheritance dependency is directional, carrying predications such as *being there* from *woman* down to *she.* But *each woman there* is just as validly the outscoper of *donkey* as *she* is, because they are identical. There is no problem inferring that *woman* outscopes *donkey.* However, Example (3) has the same pattern of scope and inheritance without such an identity (see Figure 3.2b).

(3) All the Members of Parliament assembled. Each cabinet minister held a portfolio.

In this discourse, *portfolio* is outscoped by *minister* because each minister has a unique portfolio, and *minister* inherits from *Member of Parliament* because all the cabinet ministers are MPs. But not every MP is a cabinet minister or necessarily corresponds to a portfolio. Though it seems desirable to infer that *woman* outscopes *donkey,* the analogous scope of *Member of Parliament* over *portfolio* has no intuitive interpretation.

However, formal interpretation shows that this scoping is harmless. Its truth-conditional effect can be understood in terms of the algorithm for translating a cued-association structure to a lambda term, which is given in Schuler and Wheeler (2014). In this algorithm, lambda terms are built from the inside out; elementary predications are translated first, and are then wrapped successively in variable bindings and generalized quantifiers. Preconditions on the translation of bindings and quantifiers ensure a well-formed formula.

Scope dependencies merely impose one such precondition. They delay the translation of an

outscoping property until the outscoped property is translated. Thus, for example, the scope dependency from *portfolio* up to *minister* ensures that the portfolio variable will be bound closer to the predication *hold* than the minister variable is.

The inheritance dependency also imposes a precondition on binding the minister variable. It says that whatever is true of the MPs is true of the ministers in question. Thus the nuclear scope property *being an MP that assembled* must be translated into a lambda term first, so that we can conjoin it with the property *being a Cabinet minister* to form the restrictor property for the translation of *each.* Only then can we write a lambda term for the nuclear scope of *each* that incorporates the previously translated portfolio properties.

What, then, of the inferred scope arc between *MP* and *portfolio*? Without it, both have to be bound before (i.e. within) *minister.* With it, *portfolio* additionally has to be bound within *MP,* which does not conflict with the other requirements, so it will not make the lambda translation impossible. And although scope arcs affect translation order, they cannot rewrite elementary predications, so it is still always a minister who holds said minister's portfolio, and not some arbitrary other MP.

It would be dangerous to propagate outscoper-hood up inheritance chains if entities at higher scopes could inherit from entities at lower scopes. This might cause ordering paradoxes in the translation. But an inheritance arc pointing to a lower scope is analogous to the subscripts in Example (4).

(4) *She$_i$ approved of everyone's mother$_i$.

X's mother is only meaningful under the scope of *everyone,* where *X* refers to someone. Where *she* is, *mother* does not refer, so there are no facts about *mother*'s referent that *she* can inherit. The same problem afflicts any attempt to inherit from a lower scope, making all such inheritance arcs meaningless. Since they are meaningless, they are not expected in the corpus, and ordering paradoxes cease to threaten.[6]

[6]But my preprocessor has been made robust to cycles, just in case.

3.2.1.1.1 Every woman's donkey revisited Our device for preserving common-ground information down inheritances is too simple to be accurate in anaphoric references to quantified nouns, which can be illustrated with Example (2).

Brasoveanu (2010) introduces Plural Compositional DRT in order to support correct anaphora in sentences like this one. In PCDRT, the information states that successive utterances modify are sets of variable assignments, each compatible with what has been uttered so far. Each variable assignment represents a correspondence between quantified-over individuals. Update operators in subsequent discourse must retain a structured subset of these assignments, which is to say they must preserve the correlations between individuals established by previous quantifications. In this way, when 'always' quantifies over my visits, 'she' does not quantify afresh over all the women, but only has access to the woman corresponding to my visit.

On our inheritance model of anaphora, this is difficult. The occasions quantified over by 'always' have to inherit the constraints predicated of their antecedents, such as being a visit and being by me, into their restrictor sets. The facts just mentioned are easy to capture, but 'every woman there' has to be existentialized as 'there was a woman there', to be outscoped not by the nuclear scope set of 'always' but by its restrictor set, in order to finish capturing the facts about the visit. And then, in order to refer to that visit's woman, 'she' must inherit from that virtual woman, and not (as we annotate it) from the woman variable in the previous clause.

We have explained this away among ourselves with the notion that 'always' is a disguised and displaced quantifier of 'she', so that the second clause in fact means 'every woman whom I visited had a well-fed donkey' rather than quantifying over visits. That implies and is implied by 'every visit was one with a woman who had a well-fed donkey' so strongly as to seem like the meaning of the sentence itself. Unfortunately, even if accepted in this case, that trick does not save similar situations elsewhere.

Suppose that my discourse about well-fed donkeys continued,

(5) Most of them had free access to her root cellar.

Structured anaphora would likewise require that most of the donkeys had access to the owning woman's root cellar, not to some arbitrary woman of the set, let alone to 'every' one per the quantification that introduced them to the discourse.[7] Once again, we annotate from 'her' back to 'she', and no matter whether 'she' is subordinate to a quantification over all visits or is quantified by 'always' directly, in either case it is predicated that she has a well-fed donkey. Thus the existentializing copy operation once again places a virtual 'there is a donkey that she has' into the restrictor of 'her'. However, nothing at this point forces it to be identical to whichever donkey is, at any given time, bound by 'most of those'. The correspondence has been severed.

'Most of those' has to scope high, because it ranges across all the donkeys I met, not across all the donkeys each woman owns, so its binding cannot be controlled by looking up the scope chain to learn which woman is under consideration. 'Those' as such has its own virtual copy of the preceding common ground, which includes 'there exists a woman who owns it' provided that the woman was in scope for the original 'well-fed donkey' variable.[8] But once again, we have annotated 'her' as inheriting from 'she', when what we actually need is to fetch the virtual woman from 'those'. That variable's reference can be resolved relative to the outscoping donkey. But nothing in our annotation indicates this course of action.

One argument is that nothing should have to. Plain lexical semantics and inheritance provide information supporting a pragmatic inference that the woman–donkey correspondence of the first sentence ought to be imported into the second, in preference to unlikely scenarios in which different donkeys are in the woman's vicinity when I discuss their food source than when I mention their being well-fed.

However, I and all of our annotators have been susceptible to misusing scope in an attempt to record these sorts of correspondences among individuals (see Section 3.4.2.2), suggesting that deep

[7] Of course, the problem has a different, blunter solution at this point for readers whose grammar does not allow that use of 'her', but we all would accommodate and understand it even if we wouldn't produce it, so I believe we still have to account for it.

[8] This is guaranteed by the displaced-quantifier analysis 'every woman I visited', but the 'for each visit' analysis can be modeled in ways that do or don't provide it, depending on the quantificational force we assign to the pronoun 'she', without affecting its truth conditions.

down, none of us really buy this argument of mine. Structured anaphora of some kind would be the correct annotation for what we meant.

Brasoveanu's particular approach is difficult to implement in the dependency semantics that lie between the annotation and our processed lambda expressions.[9] Jeffery King's (2004) Context Dependent Quantifiers offer 'minimal situations', which are a more natural fit and would likewise serve to relay the woman-visit-donkey correspondences from one quantification to the next.

The alternative is to devise some notation for marking an inheritance to be diverted from its literal textual antecedent to an existentialized copy of it held by one of the outscopers of the anaphor. The fact that there may be multiple such outscopers makes this particularly messy. With semantics for minimal situations, it may just be possible to solve these proportion problems while retaining our annotator-friendly surface layer.

3.2.1.2 Top-scopedness

In any document, there is at least one *topmost* scope-bearer, one without an outscoper. There may be more than one topmost, as in Example (6), so long as the topmosts cannot interact scopally.

(6) A cat meowed and a horse whinnied.

The relative scoping of two existentials cannot affect truth conditions, so in this case, both are considered topmost and no outscoper is annotated for either. We presume that topmosts such as these fall directly under a global scope. Therefore all other scope-bearers fall *indirectly* under the global scope.

However, despite transitivity, we do not supply an arc to the global scope from each and every scope-bearer. Only when a scope-bearer is topmost does its arc to the global scope reflect an annotator decision about which there could be disagreement. Such arcs we create. Arcs from other scope-bearers to global would signify nothing, and so are omitted.

[9] We employ dependency semantics for their compositionality, as described in Schuler and Wheeler (2014), and for cognitive modeling reasons beyond the scope of this thesis, no pun intended, such as incrementality and a plausible neural implementation.

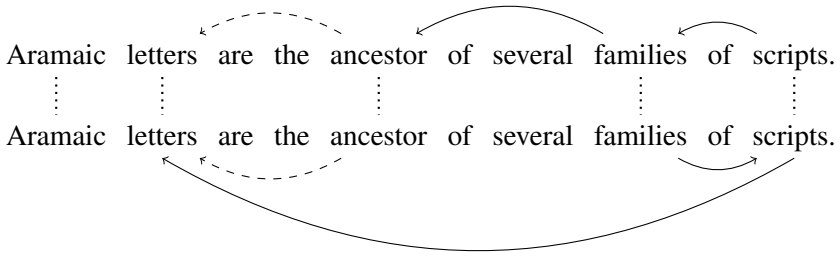

Figure 3.3: Matching two annotations of a single document.

3.2.2 Vertex Correspondence

To measure the distance between two annotations, after preprocessing, they were brought into correspondence as follows:

1. Scope-bearers correspond if they arise from the same word in the same position of the same sentence.

2. Remaining scope-bearers correspond by the order in which they appear in their documents.

3. If one document has more annotated scope-bearers than the other, top-scoped[10] dummy entries fill out the correspondence.

Under this rule, two annotations of the same document will largely be matched mention-for-mention, as shown in Figure 3.3, and a scope-bearer that only one annotation identifies as such (a 'one-sided annotation') will be represented in the other annotation by a dummy scope-bearer with an arc to the global scope and no other scoping.[11]

When comparing annotations of two different documents in the permutation test, scope-bearers are matched up arbitrarily, as shown in Figure 3.4, to provide a model of random annotations. The arbitrary matching can be imagined as adapting the structure of one scope DAG to the scope-bearers of the other (but no matter which one is adapted, the distance measure comes out the same).

[10] This provision is no longer appropriate; non-interacting dummies should be used instead.

[11] Except in the extremely unlikely case where *both* annotators provide distinct one-sided annotations.

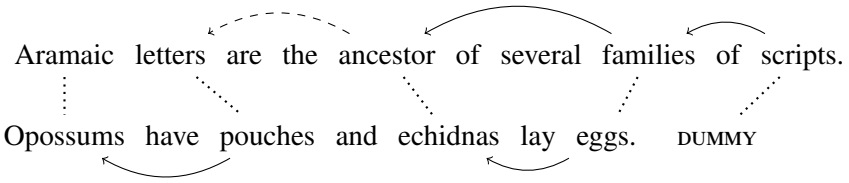

Figure 3.4: Matching annotations of two different documents.

One can imagine creating a correspondence not by this arbitrary matching but by creating dummies prolifically—in effect, treating every scope-bearer in either document as a one-sided annotation. But this is an unreasonable model of a random annotator; it implies annotators who from time to time just skip a document entirely. Arbitrary matching simulates an annotator whose product has no rational connection to the contents of the document, but who does mark scoping and inheritance dependencies just as often as the real annotators do, and who shares their slight bias toward topmost scoping.

3.2.3 Symmetric Difference as Graph Distance

A distance metric for directed acyclic graphs was introduced by Critchlow (2012). Critchlow's metric is computationally expensive (Brandenburg et al., 2012), but Malmi et al. (2015) prove an efficient approximation. With parameters appropriate for comparing scope annotations, the approximation is equivalent to the size of the symmetric difference between two graphs' sets of scope arcs. Symmetric difference is itself a metric. The remainder of Section 3.2.3 defines the metrics and the approximation, justifies the parameters, and shows how the approximation reduces to symmetric difference.

Given a correspondence between two annotations' scope-bearers, measuring the distance between the DAGs is equivalent to measuring the distance between two partial orders over the set of scope-bearers.

The conventional distance metric on total orders is Kendall's τ, defined as total pairwise disagreements (Kendall, 1938). The generalization to partial orders by Critchlow (2012) is Hausdorff

Kendall τ distance. Critchlow represents each partial order as the set of total orders that extend it; in this representation they can be compared with Hausdorff distance, which is the largest (Kendall τ) distance from any member of either set to its nearest neighbor in the other set. Because the underlying Kendall τ is a metric, the Hausdorff distance derived from it is a metric also (Hausdorff, 2005).

Hausdorff Kendall τ distance requires enumerating all pairs of total orders that extend the pair of partial orders, and all pairs of arcs in each pair of total orders. The Malmi et al. (2015) approximation enumerates all possible arcs just once, summing the following elementary distances (where $1 \geq p \geq q \geq 0$):

1 for agreeing that the arc is present and disagreeing on its direction

p for disagreeing whether the arc is present

q for agreeing that the arc is absent

0 for agreeing on the arc's presence and direction

Because disagreeing on the existence of scope arc is a disagreement about truth conditions (see Section 3.1.1), we maximize the penalty for it: $p = 1$. Likewise, agreeing on *no interaction* is an agreement about truth conditions, and we treat it just as positively as agreeing on the direction of an interaction: $q = 0$.[12]

In our transitively closed graphs, an arc is present for every scopal interaction, so the distance function amounts to scoring 1 for every arc not shared. Taking each graph as the set of its arcs, the distance is the size of the graphs' symmetric difference (their union without their intersection). Although the Malmi et al. (2015) approximation does not always satisfy the metric axioms, the size of symmetric distance does.

Summing over all possible edges and scoring disagreements is the very measure from Manshadi et al. (2011) that I previously rejected for violating independence assumptions. The difference is that now it is wrapped in a permutation test. The non-independence of pairwise comparisons inflates the

[12]Positive q increases the distance between sparse graphs. It is suggested for applications where agreeing on many arcs' presence should be favored over agreeing on many arcs' absence (Malmi et al., 2015), i.e. when an arc's presence is more meaningful than its absence. Positive q does violate two of the metric axioms (footnoted in Section 3.2).

agreement between *any* two annotations, so the random annotator reveals the extent of the inflation. Agreement calculations with and without this correction are demonstrated in Section 3.2.4.

3.2.3.1 Normalized distance functions: An open question

We would also have a metric, Jaccard distance, if we normalized the symmetric distance by dividing out the size of the union. Normalization increases the impact of an arc agreement or disagreement when the DAGs compared are small.

Our short documents usually produced small scope DAGs, often consisting of just two or three scope-bearers. As usual, in-situ scoping was much more frequent than inverse scoping, which allowed small DAGs from different documents to agree often when their scope-bearers were put into correspondence in the permutation test. Furthermore, comparisons of two small DAGs made up a much larger fraction of expected agreement than of observed agreement.

So in our case, the effect of normalization was to drive α down by an order of magnitude. Without normalization, α remained on the same order of magnitude as the κ values, confirming that the much lower result with normalization was pathological. But this in no way settles the question of whether to normalize comparisons for DAGs in general, or even for scope DAGs in general.

Skjærholt (2014) studies the analogous question for tree edit distance and concludes that the non-normalized metric is 'clearly the best'—but on the grounds that in his experiments, the normalized metric led to α scores that were modestly *higher*. And to me, his results do not make its superiority all that clear.

In simulations where he generates a second annotation by permuting the structure of gold trees to varying degrees, the normalized metric tracks closer to the current standard measure, labeled attachment score, except for the very most severe permutations, when the probability of reassigning a token to another head at random is 0.8 or above (see Skjærholt's Figures 3 and 5). And on natural data (four corpora divided into nine parts), agreement scores from the normalized metric correlate to LAS better than those from the non-normalized metric (Pearson's r of 0.5126 versus 0.4788, calculated from Skjærholt's Table 2).

But if it is not clear to me that normalizing is inappropriate for tree edit distance, neither is it a given that the best practices for trees would generalize to non-tree DAGs. Choosing good distance functions for structured annotations is a craft in its infancy.

3.2.4 Non-independence Revisited

We share with Manshadi et al. (2011) the practice of scoring the labels for every possible pair of scope-bearers, in order to extract all of the information they contain. But when they are generated to satisfy global properties like acyclicity, whether by a human annotator or an algorithm, most of them contain redundant information, and counting it multiple times overstates two annotations' agreement.

The response in Manshadi and Allen (2011) is to define additional measures that only score the transitive reduction of the scope DAG,[13] but how to combine these with or weigh them against the measures based on the entire transitive closure is not clear.

Krippendorff's α by contrast permits the distance function to overstate agreement this way in the observed-disagreement term. But in the expected-disagreement term, the distances between unrelated annotations as measured by the same function serve to characterize its overstating tendency.

By way of example, I will first illustrate agreement/disagreement arising from paired annotations of the document excerpted from the article 'Metaphor', and then chance agreement/disagreement from one annotation of 'Metaphor' and one annotation of 'Leap year'. The excerpts themselves are given in Examples (7) and (8).

(7) Metaphor is a term for a figure of speech. It does not use a word in its basic literal sense. Instead, it uses a word in a kind of comparison.

Agreement between annotations has two components: Which scope-bearers are at the topmost scope, and which pairs of scope-bearers have a scopal interaction.

[13] The unique smallest set of arcs having the same transitive closure.

In ‘Metaphor’, nine scope-bearers were identified as topmost by at least one annotator. Out of these nine possible agreements, eight scope-bearers were identified as topmost by both annotators. All eight were noun phrases. The ninth scope-bearer was a point of disagreement: One annotator also left the word *not* topmost, but the other placed it under the scope of one of the noun phrases.

Five more noun phrases were marked with outscopers by both annotators, making 14 scope-bearers in total and $\binom{14}{2} = 91$ possible pairs. Annotators agree that 84 pairs have no scopal interaction; the number is so high because of the large number of top-scoped existential quantifiers, which are incapable of interaction. In five more pairs, the annotators agree about the presence and direction of a scope interaction, and in two pairs they disagree about the presence of one. Therefore interaction contributes 89 actual agreements out of 91 possible. Together with the topmostness agreements, the total is 97 agreements out of 100 possible.

Chance agreement will be calculated the same way, except for the additional step of creating a correspondence between the scope-bearers of ‘Metaphor’ and those of ‘Leap year’.

(8) A leap year comes once every four years. It is a year in which an extra day is added to the Gregorian calendar which is used by most of the world. An ordinary year has 365 days.

An annotator found 12 scope-bearers in ‘Leap year’ and 13 in ‘Metaphor’.[14] A dummy scope-bearer was created in ‘Leap year’ to make their numbers equal.

When the two lists of scope-bearers were matched in order, there were 12 possible agreements about top-scoping. Five topmosts in each list were matched to topmosts in the other and contributed agreements. Seven more scope-bearers were marked as topmost in only one list, and contributed disagreements: three topmosts in ‘Metaphor’ and four topmosts in ‘Leap year’.

Among 13 scope-bearers, there are $\binom{13}{2} = 78$ possibly interacting pairs. Seven pairs had an arc in ‘Metaphor’, but no matching arc in ‘Leap year’, and four arcs in ‘Leap year’ had no match in

[14]The 14th scope-bearer, *not,* had no scope interactions in this annotation, and so my agreement system was unable to detect it as a scope-bearer. Had it been detected, it would have been matched with a dummy, and there would have been additional agreements about *not* and the dummy being top-scoped, and about neither one interacting with anything.

	outscoped	topmost
outscoped	134	64
topmost	86	552

Table 3.2: Topmostness confusion matrix.

'Metaphor', for 11 disagreements about interaction out of 78 possible. For the other 67 pairs, both annotations agreed that there was no arc.

Adding interaction to topmostness, total agreement in this chance matchup is 72 agreements out of 90 possible. Naïvely, 80% agreement sounds rather good! But the only thing these annotations have in common is that they were done by the same hand, and probably about the same time. This is the context in which we should understand the 97% agreement about 'Metaphor'.

3.3 Findings and Chance-corrected Agreement

Our confusion matrix for the labels of pairs of scope-bearers was already given in Table 3.1 (see Section 3.1.2). Each scope-bearer was also checked for topmostness, with the results shown in Table 3.2. The comparisons underlying both tables are only those between different annotations of the same document.

Summing through both tables, raw observed agreement is 11,062 agreements out of 11,638 possible, or 95.1%. Chance-corrected α = 60.9%.

3.4 Error Analysis

In the two confusion matrices, eight cells represent disagreements. I randomly selected five disagreements from each cell for review, then added the sixth and final disagreements from the downward/upward and upward/downward cells. In this discussion, the first mention of each cell will be boldfaced.

cause	count
annotator error	23
gap in guidelines	19
other	5

Table 3.3: Broad causes of the disagreements reviewed.

An observable disagreement about scopal interaction can arise from just one act or omission by an annotator, or from several acts or omissions interacting. Moreover, one such act or omission can contribute to several different observed disagreements. Reviewing the 42 selected disagreements revealed 47 contributions from underlying causes. In most cases, the cause was either annotator inattention or a difficult matter not anticipated by the annotation guidelines.

Table 3.3 gives the tally in these broad categories. Annotator errors are analyzed in Section 3.4.1, gaps in the guidelines are analyzed in Section 3.4.2, other causes are analyzed in Section 3.4.3, and a more detailed count follows as Table 3.4 in Section 3.4.4.

3.4.1 Annotator error

Annotator errors include oversights such as overlooking scope-bearers and mis-typing annotations; annotating contrary to guidelines that concern proper names and meronyms; and annotating scopes with strange truth-conditional consequences.

3.4.1.1 Oversights

Example (9) comes from the article titled ‘Right angle’. Subscripts have been added to distinguish two mentions of the word *angles.*

(9) When two lines cross each other so that all the angles$_1$ have the same size, the result is four right angles$_2$.

In one annotation, *all the angles* had a scope arc up to *two lines,* showing that the angles in question are specific to this crossing of these two lines. This makes the *lines–angles$_1$* pair a 'downward' scoping.[15]

In the other annotation, *all the angles* has no outscoper, amounting to a claim that crossed lines result in four right angles when all angles in the universe have the same size. This was an oversight, not the annotator's intended meaning. But because it leaves *angles$_1$* at the topmost scope, it creates an **outscoped/topmost** disagreement, and because it also expresses no interaction between *lines* and *angles$_1$*, it also creates a **downward/incomparable** disagreement.

Five of the disagreements selected for review are owed to annotators overlooking scope-bearers. Two more disagreements are because each annotator miswrote the token number for the target of one scope arc. A final three oversights (again, implicating both annotators) are idiosyncratic contradictions, either of the meaning of the scope dependency, or of the plain sense of the sentence.

These three subgroups are reported as three separate causes of error in Table 3.4, but all of them result from misperceiving the content of the annotator's raw materials, and all might be considered as variations on a theme. If they are all grouped together as oversights, they constitute the most prolific source of disagreements in the review (in a tie with one other source, to be discussed in Section 3.4.2.2).

To mitigate this problem, second-generation documents are presented to annotators with a placeholder on every nominal.

3.4.1.2 Proper Names

The annotators agreed that the reference of proper names is not (usually) relative to some other variable binding; thus they are not (usually) outscoped. This guideline was disregarded in one annotation of Example (10).

(10) North America is a continent in the Northern and Western hemispheres of Earth.

[15] I use *upward* and *downward* to describe the direction of scoping between mentions in the order in which they appear in the sentence. In the absence of inheritance dependencies, these correspond to *in-situ* and *inverse* scope.

Figure 3.5: Annotations of Example (11), discussed in Sections 3.4.1.3–3.4.1.5. Dashed arcs represent parser-supplied inheritance; solid arcs represent scope

The correct first-generation annotation would have had both *hemispheres* and *continent* topmost, since *hemispheres* is the head of proper names (even though Wikipedia writers have not capitalized it) and *continent* is indefinite. The erroneous annotation has *hemispheres* outscoped by *continent*, producing an **incomparable/upward** disagreement for the pair and a **topmost/outscoped** disagreement for *hemispheres,* both of which were selected for review.

Scope/coreference annotators are now explicitly instructed not to scope proper nouns under ordinary circumstances. Their scopal neutrality is now handled via a tag in the syntax-to-semantics annotation.

3.4.1.3 Part/whole

Annotators were instructed that wholes usually outscope their parts. In Example (11), from the article 'Cooking', *oven* was named as a part of *stove*. The two annotations are shown in Figure 3.5.

(11) An oven is a part of a stove that is like a box.

In Figure 3.5a the part/whole guideline is followed. In Figure 3.5b, it is violated by the *stove–part* scope arc, creating an **upward/downward** disagreement. Propagating the scope up the *part–oven* inheritance arc made the *oven–stove* pair an upward/downward disagreement also. Both of these disagreements were among those selected for review.

3.4.1.4 Identity, Not Scope

Figure 3.5a includes the strange assertion that *part* is both outscoped by *oven* and identical to it. Here the annotator misused a heuristic for scoping.

In cases like *two hands with five fingers,* an indication that *hands* outscopes *fingers* is that the referent of *fingers* depends on the referent of *hand*. In *an oven is a part,* it is trivially true that the referent of *part* depends on the referent of *oven*, but not because of scoping.

This error contributes to a **downward/upward** disagreement over the *oven–part* pair.[16] Another selected disagreement revealed the same annotator making the same mistake in one other location.

3.4.1.5 *Like*

Example (11) was also implicated in two other selected disagreements relating to the construal of *like*.

Figure 3.5a has *box* under *oven* with the sense that, for a generic oven, there exists one or another box that it is like. Figure 3.5b has the reverse, i.e. there is a certain box which the oven is like.

Not only was the *oven–box* pair itself one of the reviewed disagreements (downward/upward), but the same difference between the annotations emerged in the review in two other ways:

In Figure 3.5b, the *oven–box* arc combines with the erroneous *stove–part* arc that was discussed in Section 3.4.1.3 to create a downward/upward disagreement over the *stove–box* pair.[17] In addition, implementing the semantics of *that* introduces a scope-bearer (labeled *internal* in the illustration). This node inherits from *stove* because the parser erroneously attached *that is like a box* as a modifier of *stove,* rather than of *part.* Because it inherits from *stove,* it receives the same scoping relative to *box.* But even if the parser had attached the modifier correctly, there would still be a disagreement about the pair of the internal scope-bearer with *box,* because of the two annotations' different directions for the *oven–box* arc.

[16]The strange *oven–box–part* scope chain in Figure 3.5b makes it a downward/upward disagreement rather than a downward/incomparable, but the error in Figure 3.5a would cause a disagreement regardless.

[17]Either arc would suffice to make it a downward/incomparable disagreement; both are necessary to make it downward/upward.

One might argue that Figure 3.5b's scoping indicates not a certain physical box that is the prototype of ovens, but an ideal box. If that was the intended reading, these disagreements are less a matter of annotator error than of needing a guideline to identify the better of two plausible readings, a matter similar to disagreements about concrete and ideal units of measure (see Section 3.4.2.1).

3.4.2 Guideline Gaps

Several disagreements arose because of phenomena not addressed in the annotation guidelines—sometimes because a phenomenon had not been seen previously, sometimes because no satisfying treatment of the phenomenon had yet been found, and sometimes because the annotators did not foresee any difference of opinion about how to treat it.

This discussion of gaps in the guidelines will begin with annotators' two ways of handling units of measure, an issue that somewhat resembles their different treatments of *like* in Section 3.4.1.5. Other areas where the guidelines were inadequate follow: a frequent error in which annotators used scopes to express relationships that are better expressed with predications; treatment of non-quantifier scopal operators; certain forms of redundant annotation; the semantics of one-to-one relationships expressed by *once every;* and the semantics of *each other.*

3.4.2.1 Units of Measure

Annotators had not yet agreed on the proper handling of units of measure, as shown by the two annotations of Example (19) in Figure 3.6.

(12) A right angle is an angle with a measurement of 90 degrees.

Figure 3.6a implies that every angle measurement has its own degrees, an unsatisfying implication because a right angle contains infinitely many one-degree-wide spans, not just 90. Figure 3.6b implies that there are degrees or one degree that is shared by all angle measurements, an unsatisfying implication because such a degree must be omnipresent, but still of a specific, limited size. Lacking

Figure 3.6: Annotations of *angle with a measurement of 90 degrees* in Example (19).

a clear agreement about what to do here, annotators produced an upward/downward disagreement for *measurement–degrees*.

The semantics for units of measure, and how they correspond to words in the text, were subsequently agreed on (Rasmussen and Schuler, 2020), and corresponding advice added to the scoping guidelines.

3.4.2.2 Predication, Not Inter-quantifier Scope

The article 'Temple' includes the phrase *a house of worship*, which produced two of the disagreements reviewed (an upward/downward disagreement and a topmostness disagreement) because the annotators marked opposite scopes between *house* and *worship.*

Both annotators are in error. There is a correspondence between acts of worship and the places where they take place, but it is expressed by a predication: A structure exists, and worship exists, and the structure is dedicated to hosting the worship. As existentials, the two quantifications cannot interact scopally in a way that affects truth conditions.

Annotators may have been led to their errors by these factors: Scoping *worship* over *house* is a typical pattern for *of*-phrases, because most such phrases describe a part and a whole, like *the sides of the face*. Scoping *house* over *worship* reflects either the higher topicality of *house* in the discourse, or a habit of composing lambda terms with predicates in sentence order by default.

Figure 3.7: Erroneous annotations of Example (13).

If guidelines were more explicit about checking for erroneous scoping between existentials, such factors might have been less influential.

Similar errors about a sentence from the article 'Fine', shown in Example (13), produced a downward/incomparable disagreement about the *contract–rules* pair (Figure 3.7).

(13) When agreeing to a contract with a business, a customer may agree to certain rules.

Customer, *business*, and *contract* are all indefinite, and *rules* is likewise existential, so their relative scoping is irrelevant to truth conditions. The annotators' scope arcs are an attempt to bind each customer, contract, etc., into a correspondence with its counterparts from the same transaction. However, the eventualities that are predicated of them and the syntax of *when* accomplish this without additional scope.[18]

All of the disagreements in this category feature an annotator attempt to bind together existentially quantified entities by misusing scope, when in fact the entities are bound together by an eventuality in which they participate.

Errors of this kind are the largest source of disagreements seen in the review except for oversights, with which they are tied. A particular caution against this error was added to the annotation guidelines. The use of the `-w` annotation to mark unique identifiability relationships also helps to divert annotators from trying to mark them with scope.

[18] 'When' may also be analyzed as introducing a generic assertion about eventualities of agreeing to a contract, which should outscope the participants. At the time we were somewhat blind to eventuality quantification. The eventuality here is not unlike the 'minimal situation' used by King (2004) for donkey anaphora.

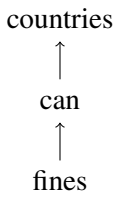

Figure 3.8: Modal *can* scoped between noun phrases in Example (14).

3.4.2.3 Other Scopal Operators

The second-largest source of disagreements seen was annotator confusion about when to annotate non-quantifier scopal operators. Annotation guidelines include 'Do not annotate by hand what can be recovered automatically from syntax' and 'Annotating non-quantifier scopal operators is obligatory only if they interact scopally with quantifiers', but *interact scopally with quantifiers* can be construed more or less broadly, and it was not spelled out which (if any) such interactions are recoverable from syntax. Six of the reviewed disagreements arose because one annotator included an operator in the scoping that the other did not. For example:

(14) In many countries, fines can be ordered by police, court judges and some government officers.

The pair *fines–can* appeared in the review. One annotator did nothing with the word *can*, whereas the other placed it above *fines* and below *countries* (see Figure 3.8), leading to an **upward/incomparable** disagreement.

The guidelines have been revised to clarify the conditions for annotating other operators.

3.4.2.4 Redundancy

Guidelines allow redundant scope annotations, but in certain cases information could be expressed either by scope or by anaphora, and there is no guideline on this form of redundancy.

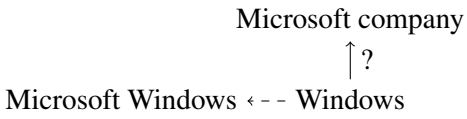

Figure 3.9: Disputed scope in Example (15). Inheritance supplied by annotators.

The pair *Windows–company,* one such case, appears in Example (15), from the article 'Microsoft Windows'.

(15) Windows is made by the Microsoft company.

Ordinarily, proper names do not need their referent annotated, but *Windows* alone occasionally refers to the X Windows system, not Microsoft Windows. As Figure 3.9 shows, both annotators noticed this and disambiguated the term by marking *Windows* in this sentence as inheriting from *Microsoft Windows,* which appeared in the previous sentence. One annotator also provided a scope arc within this sentence, from *Windows* up to *Microsoft company*. This arc resulted in an incomparable/upward disagreement on the *Windows–company* pair. The annotation guidelines do not state whether it is appropriate to redundantly disambiguate a term in this way.

Another observed disagreement concerned the topmost status of a special graph node in the semantics of Example (16), from the article 'Sport':

(16) Sportsmen need coaches to teach or train teams or individuals how to do better.

The node in question is labeled *internal* in Figure 3.10 and represents the agent of the eventuality *do better.* Embedding *how to do better* as the theme of *teach or train* makes this node inherit from *teams or individuals,* which is where the annotations actually differ.

Each annotator attempted to express the correspondence between coaches and their trainees, but not by the same means. One made *teams or individuals* anaphoric to *sportsmen* and placed *coaches* at narrow scope—incidentally leaving the internal node at topmost scope, as shown in Figure 3.10a.

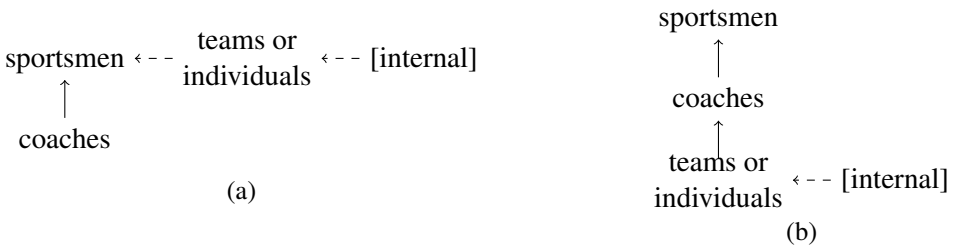

Figure 3.10: Annotations of Example (16). Inheritance from *sportsmen* supplied by annotator.

The other annotator used scope arcs to make the referent of *coaches* depend on the referent of *sportsmen*, and also to make the referent of *teams or individuals* depend on the referent of *coaches,* through the scoping shown in Figure 3.10b.

This approach establishes the coach–trainee correspondence twice, which the guidelines do not prohibit. Unfortunately, it also overlooks the anaphoric fact that sportsmen's coaches' trainees are the sportsmen themselves. And quite incidentally it makes *sportsmen* the outscoper of our internal node, creating the disagreement that was observed. Better guidelines about establishing the correspondence, such as 'prefer anaphora to scope', might have prevented the omission.

3.4.2.5 One-to-one Correspondences

One annotator used an unorthodox technique to solve a novel problem in 'Mercury (planet)' and 'Leap Year':

(17) [Mercury] makes one trip around the Sun once every 87.969 days.

(18) A leap year comes once every four years.

Example (17) means 'for each orbit, there exists an interval of 87.969 days, and for each such interval, there exists an orbit'; leap years and four-year intervals are related in much the same way.

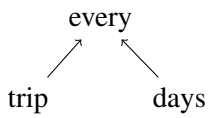

Figure 3.11: Scopal semantics of *one trip [. . .] once every 87.969 days* in Example (17).

One of the annotators recognized the semantic element that the two sentences share, shown in Equation 3.1—the assertion that two sets are in a one-to-one correspondence. But the guidelines do not explain how to annotate this assertion, and the 'prefer anaphora' guideline proposed in Section 3.4.2.4 would not be sufficient.

$$\forall p[\exists! q] \wedge \forall q[\exists! p] \tag{3.1}$$

The annotator who recognized the shared semantic element treated it as a scope-bearer in its own right and determined that it is best represented by the word *every* in its unusual position preceding another quantifier. The result was the annotation shown in Figure 3.11 for Example (17), and a corresponding annotation for the sentence about leap years.

Guidelines permit dependencies to be attached to determiners (and other non-nouns) when the corresponding noun is elided. In this case, the elided noun is something like *interval of 87.969 days*. However, the guidelines are silent on the *once every* construction specifically. Lacking such guidance, the other annotator noticed only the existential parts of its meaning, and left both common nouns topmost. The observed consequences were a topmost/outscoped disagreement on *days* and an **incomparable/downward** disagreement on *year-every.*

Semantics for intervals of recurrence have recently (autumn 2021) been worked out, allowing for guidelines about constructions like this.

3.4.2.6 *Each other*

In Example (20), repeated from Example (9) the pair *lines–each* was an upward/incomparable disagreement.

Figure 3.12: Annotations of *lines cross each other* in Example (20).

(19) When two lines cross each other so that all the angles have the same size, the result is four right angles.

The annotators agree that *each other* is in some way a reference back to *lines* but have expressed this idea differently, and the guidelines are silent.

Figure 3.12a shows *each* and *other* treated as two substantives, just as in Example (20), considering *cross each other* as a fossil of VSO word order that was permissible in the days of noun case.

(20) There are two lines. Each crosses the other.

On this view, the restrictor of *each* inherits from *lines*, and *other* is outscoped by *each* because its otherness is defined relative to *each [line]*. The noun *line* has been elided, but the guideline discussed in Section 3.4.2.5 permits the dependencies of the annotation to be attached to a determiner like *each* or a substantive adjective like *other*.

Figure 3.12b treats *each other* as a single determiner/noun phrase, whose head *other* inherits from *lines*. This treatment leaves open the question of how the reciprocal semantics of *each other* are brought about, but it avoids introducing diachronic esoterica as Figure 3.12a requires.

We have subsequently conferred, and agree that the truth conditions are approximately as found in the first analysis (we continue to debate edge cases from time to time), but we have decided that responsibility for spelling out the particular semantics of *each other* will fall to a lexicalized rule, rather than annotators building it up themselves.

Figure 3.13: Competing understandings of Example (22).

Annotators are instructed to mark scope and `-w` dependencies to the antecedent from *other*, and not to treat *each* as referential but to target later dependencies to the antecedent. This suffices to capture the information the lexical rule needs, and avoids conflict between its semantics and any others annotated by hand.

3.4.3 Other Causes

A small proportion of observed disagreements could not be ascribed to annotator error or guideline inadequacy. In most of these, both annotators seem to have a good claim to a correct reading.

Annotators disagreed about the pairs *opinions–experience* and *opinions–death* in this sentence from the article 'Soul':

(21) Many different opinions exist as to what happens to personal experience after death.

Not only this statement about opinions, but such an opinion itself must refer to *personal experience* and *death*. One annotator imagines the opinions to be generalizations, each of which introduces its own referents under conditionals. The other imagines opinions about particular actual lives and deaths. We find both readings equally persuasive.

In 'Mustache' we find this:

(22) The hair that grows on the upper lip of some men is called a mustache.

cause	count
predication, not inter-quant. scope	10
non-scopal operator	6
overlooked scope-bearer	5
part/whole	4
fair differences	3
misc. oversights	3
identity, not scope	2
like	2
one-to-one	2
proper names	2
scope-anaphor redundancy	2
typo	2
each other	1
measurement	1
software bug	2

Table 3.4: Causes of disagreements.

One annotator wanted *hair* in the scope of *lip*, as shown in Figure 3.13a, on grounds that the *lip* variable is used in the restrictor of *hair*: *the hair that grows on the upper lip*. The other annotator thought that both *lip* and *hair* fall directly under the scope of *some men*, as in Figure 3.13b, and are related to each other only through the eventuality of growing. It is a question of how *the hair* gets its definiteness, and we think the answer debatable.

Apart from fair differences of annotator opinion, one of the selected disagreements was caused solely by a software error.

3.4.4 Summary

Table 3.4 groups and ranks the 47 causes underlying the 42 disagreements that were selected for review.

A note is in order about the distribution of disagreements across documents.

The 42 selected disagreements (out of 576 total) touched on 18 of the 33 IAA documents. The most error-prone document was the excerpt from 'Cooking', both in the selection (8 out of 43;

18.6%) and in the whole IAA test set (74 out of 576; 12.8%). Both annotators did very poorly on this document.

Nearly all of the other documents represented in the selection contributed about the same fraction of its disagreements as they contributed to the whole IAA set, but by chance the selection strongly over-represented ‘Soul’ (14.0% of reviewed disagreements, only 4.7% of all), ‘Right Angle’ (7.0% of reviewed, 2.4% of all), and ‘North America’ (2 disagreements reviewed, 4.7%; versus 3 disagreements in all, 0.5%). Their over-representation comes at the expense of the documents not touched on by the selected disagreements.

Chapter 4

Task framing and data preparation

4.1 WSC as analogous task

To frame the scoping task in encode-and-classify form while treating scope as a dependency, as proposed in Section 1.4.1, I emulated the WSC task from SuperGLUE (Wang et al., 2020), as implemented in `jiant` version 2 (Phang et al., 2020).

WSC is a task of resolving pronoun references that are highly sensitive to semantic context, a superset of the original Winograd Schema Challenge (Levesque et al., 2012). Sample problems[1] show that as intended, world knowledge often points to the correct antecedent. In items like

(1) Nancy not only provided the policeman with an excellent description of the heavyset thirty-year-old prowler, but drew a rough sketch of **his** face.

the gold-standard human judgement may also have been influenced by other cues such as information structure and syntactic parallelism. The items from the original Winograd Schema Challenge come with an additional, stricter proof of context-sensitivity, a single-word alternation elsewhere in the sentence that changes human raters' preferred antecedent. So the range and importance of contextual information in WSC is somewhat analogous to the factors of scope judgement.

Previously the Winograd Schema Challenge has been framed in terms of substituting each candidate antecedent for the pronoun, and then either calculating the probability of the resulting sentence (Radford et al., 2019), or predicting whether it is entailed by the original (the WNLI framing; Wang et al., 2019). In the WSC framing, the pronoun and a candidate antecedent are just two marked

[1] `http://commonsensereasoning.org/disambiguation.html`

spans of the text, and the system must classify the relationship between them: coreferential, or not? Unlike substituting antecedents for the pronoun, feeding spans into a classifier is a general technique applicable to other long-distance dependencies, such as our scopal relations.

4.2 Item preparation

A single training or test item for scoping thus consists of a text, two spans, and (in training) a label. In WSC the two spans consist of a noun or noun phrase and a pronoun; here they consist of two nominals. The WSC labels and predictions are binary. Here, it is in question not only which but *whether* any particular scoping between the nominals is necessary to reach the intended reading of the text, so there may be three or even four labels. The data item is labeled 'direct' or 'inverse' when the intended reading does require a certain scoping, but must take some other label when it does not.

In the three-label system, used by Manshadi and Allen (2011), all other items are 'none'. The four-label system, used by Andrew and MacCartney (2004), distinguishes two kinds of 'none's: 'Equivalent' items are those where either quantifier may be in the other's scope without affecting truth conditions. 'Independent' items are those where neither quantifier need be in the other's scope, because there is no predication to which both variables are arguments.

Although the greater specificity of the four-label system is appealing, this project may have three or more quantifiers interacting in a single sentence, as Manshadi and Allen (2011) did. Their measure for scoring multi-quantifier scoping relies on the three-label system, so for compatibility, this project begins with the same.[2]

However, in preliminary experiments, validation scores from a 3-way predictor never quite beat the majority-prediction baseline (then 89.65%). I divided the labor on the same lines as Andrew and MacCartney (2004) and attempted only to predict the direction of verified scope interactions,

[2]Our corpus often has equivalent scopings among existential quantifiers from a single sentence, and ubiquitously has non-interaction of the "independent" kind between quantifiers of different sentences. Anyone interested in the four-label system can extract the independent items in their myriads from the lambda-expression form of the corpus (see Section 2.1.11).

by filtering the items to exclude "none"s.[3] This division of labor immediately proved more tractable (validation scores at that time up to 85.4% versus that data's majority-baseline 67.5%), and was pursued further.

I now also train a separate predictor for the complementary subtask of predicting whether there is a scopal interaction. Thinking in terms of an application pipeline, this predictor would screen items before sending some to the direction predictor Thus I build two reductions of the ideal ternary data down to binary: One filtered for the direction subtask, and the other retaining the "none"s but conflating "direct" and "inverse" to "scoped," for screening.

4.2.1 Extracting necessary scopes from annotations

As mentioned in Section 2.4.2.1, the annotations in our corpus are not simply a statement of all and only the truth-functionally necessary scopal interactions among nominals. Most obviously, scopings implied by transitivity are left implicit in the annotation. But before deriving these, there are other implied scopings, and scopings annotated in unusual ways, that must be collected; and afterward, there are by-products of the data pipeline that must be removed from the task set.

Each scoping emerging from this extraction process is the nucleus of one data item. Fleshing them out into the full WSC-like form, with the document text etc., is described in Section 4.2.2.

4.2.1.1 Notational complexities

The implicit scopings and unusual annotations communicate two kinds of scope interactions: many of the immediate outscopings of narrow-scoped existential quantifiers, and all necessary outscopings that do not form a tree. There is some overlap between the two kinds.

4.2.1.1.1 Implicit existential sinking The other implied scopings are those covered by a convention we call 'existential sinking'. Briefly stated, an unannotated existential quantifier implicitly

[3] This is somewhat analogous to dividing anaphora resolution into subtasks of identifying candidate antecedents and selecting among them.

takes the narrowest scope sufficient to avoid unbound variables in predications, or a scope made wider than this only by interposing other unannotated existentials.

This convention was originally created to address the existentially quantified eventualities that a Neo-Davidsonian semantics (Parsons, 1990) must create for every predication, whose scope is almost always[4] narrow as described, although it also applies to existential quantification over entities. 'Sinking' saves time annotating their scopes and allows less predictable scoping to stand out visually.

Existential sinking also tidies the corpus in another way, by sidestepping the biggest difference between two ways of using scope data, which might be called 'tree-oriented' and 'DAG-oriented'.

The other major consumer of our annotated documents is tree-oriented. It is software that compiles them out into a corpus of lambda expressions. Since lambda terms are tree-structured, it must assign exactly one immediate outscoper to each quantifier in an expression (except for the topmost). Transitive outscopings are implicit in these, and so need not be tracked directly. In the presence of equivalent scopings (as defined in Section 4.2), more than one lambda term can express the same truth conditions, and the software arbitrarily produces one of them. In such a case, the term will contain outscopings that are merely allowed by the correct reading of the text, as well as outscopings that are actually required for it. The software has no need to track which is which.

This project is DAG-oriented. So that scope predictions can be scored fairly, it must not assign the same status to allowed as to required outscopings, and it must track the required ones, regardless of whether they are immediate in every lambda term with the correct truth conditions, or immediate in some but transitive in others. In other words, it needs different data from the tree-oriented software precisely where there are equivalent scopings.

Where necessary, these differing requirements are met with double notation: distinct sets of tree-oriented and DAG-oriented annotation tags. This system is described in Section 4.2.1.1.2. However,

[4]In '13,674 people set the first world record for simultaneous skinny dipping', the setting-a-record should outscope the people. It is true but weak that for each of the people there exists one or another setting-a-record such that the person participated, it was the record whose count was 13,674, and so on. To assert that they all participated in the same setting-a-record, we must either put it at higher scope, or somehow bake it into lexical semantics of 'world record' that the (single, high-scoped) record they set uniquely identifies the eventuality of setting it.

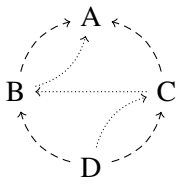

Figure 4.1: Double notation for dual-use data

the vast majority of equivalent scopings come from adjacent narrow-scoped existential quantifiers, i.e. the ones covered by the existential sinking convention.

Sinking therefore suppresses a thicket of double notation that would obscure the more interesting, less predictable annotations—at the cost of some preprocessing for everyone. Tree-oriented users must fill in immediate outscopings that are not strictly necessary, both among the sunk existentials and from the topmost of them to the next higher non-existential. DAG-oriented users can neglect outscopings among the sunk existentials entirely, and infer outscoping that are necessary but not strictly immediate from each of them directly to the non-existential.

4.2.1.1.2 Double notation As mentioned, on occasion, a document has multiple equivalent logical forms not covered by existential sinking. In such cases we use double notation. Figure 4.1 represents this schematically.

One of the equivalent alternatives is selected, and special annotation tags (dotted arrows in the figure) trace its immediate outscopings for the lambda generator. Other special tags (dashed arrows) mark any-distance outscopings that are necessary to the intended meaning, but need not be immediate, i.e. the outscopings that are immediate in some but not all of the equivalent logical forms. The regular annotation tags can still be used above and below the double-annotated region of the graph, for outscopings that are immediate in all of the logical forms.

Conceptually, we need only ignore the lambda generator's special notation, read off the scopings from our own notation, and add them to the ones from the common, shared notation. In practice, we accept some extra complexity handling these in order to reuse existing code for existential sinking.

4.2.1.2 Existential sinking in practice

The statement of the existential sinking convention in Section 4.2.1.1.1 is succinct but inexact. The actual rule has edge cases and idiosyncracies because of its development history.

Because some of the outscopings implied by sinking are truth-functionally necessary, the scope prediction project needs an implementation of existential sinking. I have reused the implementation from the lambda generator. Although its tree-oriented output has to be post-processed for this DAG-oriented application, we found this less costly than an exact reimplementation.

The lambda generator does not translate from syntax directly to lambda terms, but takes the information through an intermediate representation as dependencies. It is on this representation that it performs existential sinking and other useful processes, such as copying semantics from conjunctions down to their conjuncts.[5]

After these processes, which we collectively call normalization, the lambda terms proper can be read off of the dependency graph with the Schuler and Wheeler (2014) algorithm. I added an option to dump the normalized dependencies instead of proceeding to read-out.

This dump is not entirely suitable for the prediction project. The fact that it contains only immediate outscopings is of minor importance, because we can transitivize. However, the fact that the normalizing code works only in terms of immediate outscopings causes information to be lost (or at least buried).

Providing immediate outscopings means the normalizer must adopt outscopings that are not strictly necessary whenever two existentials sink to the same place, and in all other cases of multiple truth-conditionally equivalent logical forms. It takes its half of double notation (Section 4.2.1.1.2) as authoritative where present, breaks ties arbitrarily otherwise, and does not distinguish between the outscopings so created and the ones that are genuinely necessary to obtain the correct truth conditions. All of these are hindrances to a DAG-oriented application, but unfortunately, the process of imputing outscopers to unannotated existentials is woven into them.

[5] Since scopes may be among the semantic information annotated on a conjunction, it is necessary and proper that these processes run together.

To obtain the results of existential sinking but also preserve other information this project needs, I therefore modified the syntax-to-semantics translator and subsequent processing code beyond just adding the dump option. The modified code passes both halves of the double notation through to the output, unchanged except for the conjunction-related processing.

The arcs for necessary but not strictly immediate outscoping could be reconstructed with algebraic manipulation of the lambda terms, but passing them through normalization spares us the trouble. The arcs for immediate but not strictly necessary outscoping are duplicated when the annotations are read, and the processing code is allowed to continue taking one copy as authoritative, so that its work will proceed exactly as before. The other copy is passed through separately, to preserve the information that the outscoping is arbitrary (or, colloquially, to mark it as tainted).

Helpful as this is when there *is* double notation, for existential sinking it does nothing. Further processing downstream would be necessary.[6]

4.2.1.3 Preparing clean scopings

The data preparation pipeline thus began with the annotated corpus files and reached the desired scopings through three programs and a small amount of hand-hacking.

The syntax-to-semantics program produced 'discourse graphs' (semantic dependencies). A small number of annotated documents contained syntactic oddities not yet accommodated by this program, and the resulting error messages had to be hand-culled from its output. A smaller subset of these documents had errors that could not be worked around, and did not yield discourse graphs at all.

The modified normalizer performed the desired existential sinking on the discourse graphs. Another small subset of documents was lost at this stage, this time because of semantic oddities that could not pass the lambda generator.[7]

[6] It appears to me at present that the further processing in fact makes the passed-through arcs unnecessary. At the time I designed the data processing pipeline, I believed I knew a possible scenario where they were indispensable. But if the passed-through arcs are redundant, at least they did not corrupt the data, were quick to implement, and made it easier to reason about the rest of the problem.

[7] Some annotation errors were caught and fixed this way.

A postprocessor, described in this section, winnowed out the desired information about truth conditionally necessary scopings and formed up the list of truth-conditionally meaningful any-distance outscopings.

4.2.1.3.1 Sift and untaint Working document by document, the postprocessor first sorted the dependency arcs of the normalized graph. It discarded arcs for predicate-argument semantics and built collections of the arcs that mark immediate outscoping, any-distance necessary outscoping, and truth-functionally arbitrary ('tainted') outscoping. For use in later processing, it built lists of the graph nodes representing existential and universal quantifications, and of the arcs joining these to their restrictor and nuclear scope sets.

The lists of quantifications collected in this step have one peculiarity, connected with the article 'the'. When definite descriptions are considered as referential, most of them denote a single entity (and have singular grammatical number). The equivalent, when they are considered as quantificational, is that their restrictor set is a singleton.[8] And although I understand a definite article's asserted meaning as a universal quantifier (as discussed in Section 5.1.3.8), when the restrictor set is a singleton, an existential has exactly the same truth conditions: If there is exactly one Dalai Lama, then whenever some Dalai Lama sneezes, every Dalai Lama sneezes (and, of course, *vice versa*). The upshot is that many noun phrases with 'the' are logically indistinguishable from existentials.

Moreover, we frequently find these definites at the narrowest of scopes, adjacent to the sunken existentials. In this position, their equivalence to an existential includes them in the existentials' freedom to nest scopes in any order without affecting truth conditions. The normalized dependencies make it very easy to identify definites of this kind in this position, so they are included in the list of the graph's existential quantifiers.[9]

[8] A singleton within its scopal environment. That is, 'Each bus stopped and the driver got out' may describe five buses and five drivers, but when considering any given bus, 'the driver' is a single individual.

[9] Strictly speaking, they should also have been included in the list of universal quantifiers, but true universals (as opposed to generics) are much rarer in the data than existentials, so this is unlikely to have made much difference.

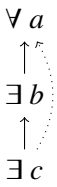

Figure 4.2: An irrelevant arc is needed to infer a relevant one

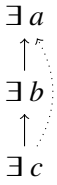

Figure 4.3: An irrelevant arc (briefly) comes into being

Immediately after this step, the list of immediate outscoping arcs was filtered to eliminate those that match a taint-marker arc. At this point it was not yet safe to eliminate all arcs without truth-conditional effect, for reasons illustrated in Figure 4.2. The arc from ∃ *c* up to ∃ *b* is such an arc, but without it, the arc from ∃ *c* up to ∀ *a* cannot be inferred. Taint markers, however, arise from double notation, which also supplies the truth-conditionally necessary non-immediate arcs that detour around the tainted ones, so that the problem does not arise.

4.2.1.3.2 Double transitivization Taking the transitive closure of the immediate arcs will at this point produce all of the outscopings that are to become data items, but it will also produce more of the arcs we wish to eliminate, as shown in Figure 4.3. The relative scope of two existential quantifiers is without truth-conditional effect not only when one immediately outscopes the other, but at any distance, so long as only existentials intervene.

From this it follows that whenever it *is* truth-conditionally necessary for one existential to outscope another, some non-existential quantifier must fall between them. The arcs from the lower existential to the non-existential, and from there to the higher existential, are themselves truth-conditionally necessary. They also transitively imply the arc between the existentials.

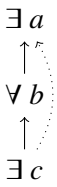

Figure 4.4: A necessary arc is destroyed and re-created

This provides a convenient way to finally eliminate unnecessary outscoping arcs from the data. Wherever a contiguous chain of one or more existential quantifiers outscopes or is outscoped by any non-existential, by taking the transitive closure a first time, among other arcs we create truth-conditionally necessary arcs directly between that non-existential and every existential in the chain.

Now we can drop *every* arc whose endpoints are both existentials. By so doing we lose all of the unnecessary outscopings within the contiguous chain. We also lose necessary outscopings that cross from one contiguous chain to another over a non-existential. However, after the first transitivization, each of these necessary arcs is in a position similar to Figure 4.4. It can be reconstructed just by taking the transitive closure a second time.

The lists built and set aside when sorting the arcs were used in this step. The discourse graph marks scoping on the sets/properties that are quantifiers' arguments, but marks quantificational force on nodes representing the quantifiers themselves, so identifying scope arcs between two existentials requires some data to bring them together.

Outscopings within a contiguous chain of universal quantifiers are without truth-conditional effect also. *Mutatis mutandis,* the same reasoning applies, and the same two transitivizations can prepare and repair both kinds of arc-dropping. I implemented both.

The final remnant of double notation, the necessary upward arcs, can be added into the data before either transitivization. Since the first transitivization can do what it needs to without them, I held them until the second.

4.2.1.3.3 Scopings not attempted After these steps, the known scopings include scope-bearers and scoping patterns beyond the needs, ambit, or abilities of this project. Up through transitivization, these had to be retained to avoid information loss, but keeping them around longer would entail complications not desirable for a limited, well-defined attempt at predicting scope with a new technology.

The known scopings include quantifications that cannot be straightforwardly construed as having a one-to-one correspondence with a word or phrase. It is difficult to reconcile these with a prediction approach that revolves around classifying the relationship between two spans of text.

The known scopings include quantifications nested not only within other quantifications' nuclear scope sets, but within their restrictor sets, such as when restrictive relative clauses contain quantification:

(2) Most netizens know someone that doesn't have a cat.

In fact, this is a special case of the one-to-one problem; the word 'someone' can stand for a nuclear scope set (which is the object of a netizen's acquaintance), or for a restrictor set (which immediately outscopes a negation), but making it represent both at different times is a complication better deferred. The most common and familiar examples of scopal interaction involve nuclear scope sets only, so for purposes of the present work, I construed these rather than restrictor sets as the referents of nouns and pronouns.

Finally, the known scopings include quantified eventualities and non-nominal operators (such as the negation above, or modal verbs). But again, these are less prototypical scopal interactions than those of nominals, and less common, so I chose to exclude them from early work.[10]

Present purposes therefore required filtering the known scopings. The final step in the post-processor just described was to begin this filtering. It took advantage of systematically constructed node IDs in the discourse graph to identify nuclear scope sets directly denoted by a single word, and to keep only those scopings whose endpoints both met this description.

[10]Expanding to cover these operators is a logical future step.

Other programs downstream excluded modal verbs and eventualities by verifying that the words in question were nouns or pronouns. I also used them to eliminate a huge, almost entirely trivial source of 'no interaction' items by dropping scopings that crossed sentence boundaries, after a pilot project showed that these outnumbered interesting scoping items by orders of magnitude.

4.2.2 Rendering items

A complete data item, presented as a span pair classification problem in the image of the Winograd Schema Challenge, requires information about the document that is lost in the foregoing semantic processing: its full text, the texts of the two spans, and their positions within the document. In addition, expanding an outscoping to a complete data item requires reconciling some technical differences between the ways of our corpus and the ways of RoBERTa, the language model that encodes the spans for classification. Finally, a complete *set* of data items for a document includes an item for every pair of nominals that are in the same sentence, including the 'no scopal interaction' items. Text processing cleared these hurdles in two steps—first collating the necessary information from the annotated corpus files, then using it with the extracted outscopings to generate and write out ready-to-use data items.

4.2.2.1 Collating the formalities

The first step went back to the corpus files to recover, for each document, its tokenized full text and a table of all its nouns and pronouns. The table contained their positions in two formats, one used in the corpus and the other in the RoBERTa tokenizer.

The corpus treats texts hierarchically, as a list of sentences which consist in turn of tokens. This lends itself to a hierarchical description of token positions: sentence number, then position-in-sentence. It also delimits each document's title from its first sentence as a matter of course. RoBERTa uses a flat list of tokens, and positions are described accordingly).

The software for this step therefore maintained running counts in both forms as it extracted tokens from each document's sentences' parse trees. It used these counts to build a new table entry

```
{"idx": 306, "label": "direct", "target": {"span1_index": 26,
"span1_text": "human", "span2_index": 32, "span2_text": "nose"},
"text": "1 ( number ) . One ( 1 ) is a natural number after zero and
before two . It represents a single item . A human typically has one
head , nose , mouth , and navel ( belly - button ) ."}
```


Figure 4.5: A `jiant`-ready scoping item in JSON format.

whenever the token's preterminal category was nominal. It also inserted a separator at the end of each document's title.

After processing each document in this fashion, I culled the few that had not survived the normalization pipeline, and pasted the rest onto the corresponding lists of extracted scopes.

4.2.2.2 Writing out data items

When writing out complete data items, the tokenized text is used in full (as input to the encoder) and to generate the two spans.[11] The table of nominals has numerous uses.

After reading in the extracted scopings, the table is used to drop those that use a non-nominal, and to translate the token positions of the others into the flat format. Comparing the positions of outscoper and outscoped determines whether the scoping is direct or inverse, and this judgement is recorded.

The table is then used to generate pairs of nominals, with the constraint that the two must come from the same sentence. These pairs drive the remainder of the process; each is checked against the recorded judgements to determine its gold label, and the entire data item is built as a JSON string. Figure 4.5 shows an example.

Finally, items in training sets are shuffled, each item is assigned a unique identifier, and the two reductions of the data are prepared: the filtered subset with scopal interactions, for the direction task, and the version conflating the two directions, for the screening task.

[11] They are provided as text, not just as indices into the text, so that differences between the data's tokenization and the encoder's tokenization can be smoothed over.

4.2.3 Train/val/test splits

I divide the data into training, development/validation, and test sets at the granularity of documents, not of individual items, in the common 8:1:1 proportion. Different documents generate different numbers of data items, so at that granularity the proportion is only approximate.

I use document granularity because any given word token is encoded the same in every data item where it is found.[12] We fully expect this encoding to reflect several qualities that predict scoping: choice of quantifier, syntactic environment, and the properties of the entity the nominal represents.

There may be as many such items as there are other nominals in its sentence, and each of these other nominals may likewise participate in several data items. When a document creates such a family of items, training on nearly all of them and then testing on one may not literally be peeking at the test during training, but it comes uncomfortably close.

In the worst possible case, it risks giving the appearance of great success to a model that memorizes a nearly complete scoping graph for each document during training, and acquires a minimum of generalizable knowledge (perhaps nothing more than transitivity) sufficient to infer an arc or two after retrieving that graph from memory.

For a less extreme, more probable, and still undesirable scenario, consider a sentence with three nominals in direct scoping, forming a transitive triple, as shown with arbitrary labels in Figure 4.6. If the training set contains the two shorter arcs, CB and BA, they tune the encoder and classifier for words A and C in their respective contexts, to produce a representation of word A that inclines the classifier to predict direct scope when A is in the left position, and a representation of the word C that has the same effect when it is on the right.

If arc CA is in the test set, the test item will contain the very same representations of A and C in these very positions. That literal item may not be in the training set, but it is the most straightforward possible hybrid of two items that were, and not a very rigorous test. Many other common scoping patterns present variations on this theme of reused representations, but the issue is avoided when training and test items are from entirely separate documents.

[12] Except that it may be the head or the tail when the two words' encodings are concatenated, depending on word order.

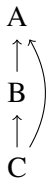

Figure 4.6: A transitive triple leads to reused word representations

A final consideration when splitting the data is that different parts of the corpus are the work of different annotators, each of whom ascended a learning curve. I distribute documents round-robin to ten folds, to avoid a test or validation set specially enriched in work by one person or at one proficiency level.

For the screening subtask, one fold is arbitrarily taken for the validation set, another for the test set, and the rest for training. The documents are then rendered down to 22,250 individual scoping items by a process detailed in Sections 4.2.1–4.2.2).

Unfortunately, filtering for the direction subtask leaves only 588 items, and such a small dataset is vulnerable to sampling artifacts, so I cross-validate this subtask; the validation set comes from the same fold as before, and each of the other nine folds serves once as the test set with a classifier trained on the other eight. This allows 523 items to be tested, instead of just 60 or so from a single fold.

4.2.4 Summary of prepared data items

Table 4.1 summarizes the data items produced by the pipeline. Fold 1 received 52 documents; all other folds received 51. The screening task used fold 1 as its test set (2226 items, of which 67 had a scopal interaction) and folds 2–9 for training (17,837 items, of which 456 had a scopal interaction).

Fold	No interaction	Direct scope	Inverse scope
1	2159	38	29
2	2027	20	12
3	2103	48	20
4	2245	33	14
5	2112	48	14
6	2284	59	28
7	2116	47	12
8	2368	43	15
9	2126	32	11
Validation	2122	48	17

Table 4.1: Summary of data items

4.3 Use of `jiant`

I trained my models using `jiant` (Phang et al., 2020) v2.1.2, a library designed to use a large pretrained language model from HuggingFace Transformers (Wolf et al., 2020) as an encoder for one or several *task heads,* which may be linear classifiers, span extractors, masked token predictors, etc., for experiments in transfer learning.

It is built for downloading shared tasks and uploading predictions to a leaderboard, but it accommodates locally prepared data, and the code that formats predictions for SuperGLUE (Wang et al., 2020) can be cannibalized to extract predictions for local viewing and analysis.[13]

4.3.1 RoBERTa as encoder

Having established that contextualized word embeddings generally capture the factors that correlate with scoping (see Section 1.3.1.1), and that among encoders for these, large attention-based pretrained language models do learn highly context-sensitive long-distance semantic dependencies in various tasks (see Section 1.3), I needed to choose such a model.

[13]The code that scores validation sets during training allows many sophisticated kinds of evaluation against gold labels and could, with some effort, be abused to score test sets as well. For my simple scoring on accuracy, this was excessive.

Much of the relevant research had used BERT (Devlin et al., 2018), which would have been a reasonable choice, but I found that RoBERTa (Liu et al., 2019) was also available. It is a variation on BERT, with an improved Byte Pair Encoder (Sennrich et al., 2016) and a few other simple changes to tokenization and pretraining. These allowed it to improve on BERT's performance for a wide variety of tasks. Many other models were and are available, but after discounting those more radically different from BERT (trained on other or multiple languages, on multimodal data, with entirely new pretraining objectives, etc.), RoBERTa stood out as a model likely to share the properties that make BERT a plausible scope predictor, but likely to outperform it.

RoBERTa is available in two sizes. To improve training speed, I worked with the smaller one: 125M parameters, with 12 layers, 768 nodes per hidden layer, and 12 attention heads.[14] I used default values for all configuration of the tokenizer and the model architecture, except for expanding the tokenizer's maximum sequence length from 128 tokens to 256.[15]

4.3.2 Span comparison classifier as task head

After encoding with RoBERTa, I used `jiant`'s span comparison classifier to label scopal relationships.

As the name suggests, this task head begins by calculating a number of span encodings from token encodings, using the AllenNLP self-attentive span extractor (Gardner et al., 2018; Lee et al., 2017). Although my two spans are always single words, they may be tokenized as a whole or as multiple sub-words. The span extractor serves to produce a single representation for the word in either case.

The span encodings are then concatenated in an order determined by the data item. I used only left-to-right word order. Dropout is applied, then a linear classifier. The screening subtask was built

[14]For unclear reasons, when work began, the compute nodes available to this project ran roughly an order of magnitude slower than they should have. Later, an update to the job scheduler corrected this, but it was a poor time to revisit basic architectural decisions. It would certainly have been preferable to use RoBERTa-large if that had been possible at the outset, particularly in light of Tenney et al.'s 2019 results on the Winograd task (see Section 1.3.1.1.2).

[15]Our longest document was 188 tokens. The default maximum in the model itself, unlike that of the tokenizer, was already large enough.

to purpose and uses a two-output classifier. The direction subtask retains the three-output classifier of the whole task it was adapted from.[16]

4.3.3 Hyperparameters

Learning rate, gradient clipping, hidden-layer dropout, and attention dropout values were roughed in iteratively using a preliminary version of the data, and the neighborhood thus identified was grid-searched with the final data, training on folds 2–9 and predicting the labels of the validation set (scored by accuracy). The chosen hyperparameters represent a consensus of the runs with highest validation scores in this final grid search.[17]

Iterations grid-searched at least two hyperparameters at a time, to account for non-independence.[18] All runs in an iteration used the same seed for random number generation. For each run, the best of several validation scores was taken. Trends in these scores directed the next iteration's search ranges. Matching trends in the runs' final validation scores were used to confirm that the search was improving accumulation of knowledge in the model, and not just increasing the chance that extreme updates would briefly drive a model through a useful state. Iterations continued until performance plateaued.

The final grid search was run similarly. For each combination of hyperparameters, the screening task was trained for 30 epochs, with 45 regularly spaced validations.[19] The direction task was trained for 120 epochs, validating after each. Several runs tied for highest validation score and were collated.

Where the collated values of a hyperparameter showed a clear best choice, that value was taken. Runs with these values were given additional weight, to clarify the best choices for remaining parameters. Where the group of highest-scoring runs was too small for any clear consensus, the next-highest were added.

[16] Training on two-output data sufficed to ensure it behaved as a two-output classifier at test time.
[17] Earlier iterations showed no effect of varying the Adam optimizer's epsilon, and it was left at its default 10^{-7}.
[18] For example, for best performance, learning rate had to co-vary with dropout.
[19] A default setting limited these to the first 500 items out of 2187 in the validation set.

	Screening	Direction
Learning rate	$3.16 * 10^{-5}$	$7.50 * 10^{-6}$
Max gradient norm	3.2	4.6
Attention dropout	0.08	0.20
Hidden-layer dropout	0.08	0.23

Table 4.2: Final hyperparameters by task

Table 4.2 gives the hyperparameters that were chosen. Because the grid search was trained on folds 2–9, the models with those hyperparameters could be directly reused to make predictions on fold 1/the screening task test set. Of course, new models were trained for the other folds of the direction task, again taking the model with the best validation score out of 120 epochs.

Chapter 5

Results and analysis

The next sections discuss various measures of the model's success on the direction task. The screening task was less successful and is reviewed separately in Section 5.3.

5.1 Accuracy and other counts of predictions

Models were evaluated before all else on accuracy, i.e. the fraction of all items predicted correctly. This has been the common measure of success in previous work on scope prediction, whether with exactly two scope-bearers (Higgins and Sadock, 2003; Andrew and MacCartney, 2004; Srinivasan and Yates, 2009; Dinesh et al., 2011) or generalized to any number of them (Manshadi et al., 2013).

Because our data set is transitively closed, accuracy in a full 3-way classification is identical to the Manshadi et al. σ measure. Accuracy in the screening subtask is equivalent to scoring the 3-way task with a relaxed definition of 'correct prediction' that treats the two directional predictions as interchangeable with each other, and only 'no interaction' as distinct.

The effect on accuracy of the direction subtask is more ambiguous. Excluding the 'no interaction' cases removes a source of score inflation, since this is by far the most numerous outcome, but it leaves fewer ways to mispredict the items that remain. In any case, this means that accuracy scores are not comparable across subtasks.

Because the predictor works arc by arc, without biases incurred by reconciling one prediction to another to make the graph acyclic and transitively closed, the accuracy measure can be used directly, without the correction procedures it was wrapped in for IAA.

Items	RoBERTa	Baseline
67	**77.61%**	56.72%
32	**81.25%**	62.50%
68	**75.00%**	70.59%
47	**72.34%**	70.21%
62	69.35%	**77.42%**
87	**81.61%**	67.82%
59	**79.66%**	**79.66%**
58	**82.76%**	74.14%
43	**83.72%**	74.42%
Total 523	**78.01%**	70.36%

Table 5.1: Accuracies for direction

Most previous work used accuracy alongside other measures. I do likewise in Section 5.1.2, to better characterize the performance of the current models relative to the current baseline and relative to future work.

Finally, inspired by the well-known tendency of 'each' to take outer scopes, in Section 5.1.3 I examine accuracy on particular determiners.

5.1.1 Overall accuracy

Table 5.1 gives accuracies for the cross-validated direction task, compared to the majority-prediction baseline. Because folds containing equal numbers of documents do not necessarily produce equal numbers of items, the number of items in each fold is given also.

The trained system outperforms the baseline in aggregate, and on seven folds out of nine (an eighth fold being a tie), but the test sets are small, and the variance is hard to ignore for either system. The sampling distribution for the trained system's accuracy is unknown,[1] so many customary, parametric tests of significance are inapplicable, but permutation offers a suitable nonparametric test.

[1]Cross-validation is an unbiased estimator for its center, but a high-variance one, and there is no general-purpose, unbiased estimator for that variance (Bengio and Grandvalet, 2004).

The null hypothesis, that the trained system and the baseline are equally accurate on arbitrary documents, is simulated by exchanging their predictions with 50% probability for each document. The alternative hypothesis is one-sided, i.e. a simulation supports the null if its performance gap is equal to or larger than the observed one and in the same direction.

Document granularity is used because different items involving the same word token use the same encoding of it, as discussed in Section 4.2.3). Exchanging predictions at the granularity of single items underestimates the variability caused by any given encoding being especially adequate or inadequate for the classifier's use, and thus overstates the significance of the observed accuracy gap.

After 100,000,000 simulations, $p \approx 0.004$ and the null hypothesis is rejected.

This predicts that the training data and procedure will generalize better than the majority-prediction baseline to arbitrary new Simple English Wikipedia excerpts prepared in the same way. That interpretation takes the models for the folds, each trained on a large subset of the data, as approximations of a model trained on all current data. The finding of significance also suggests that the procedure would generalize well to other training sets, but we should be cautious of overinterpreting it.

On one hand, using broad-domain data by many authors is intended to support generalization; consistent authorial habits or single-domain patterns, such as the meronymy Schuler and Wheeler (2014) exploited for QuanText, would likely make the task easier. On the other hand, the finding of significance rests on an implied estimate of variance for procedure and data. As an estimate of variance for the procedure alone, it is biased low, because the shared training data from document to document, within and across folds, makes the documents non-independent trials (Demšar, 2006).

5.1.2 Other item-counting measures

Accuracy is the measure previous scope-prediction efforts share, but most also used other measures, to give a better picture. This would help to characterize the performance of the current models relative to the current baseline and relative to future work.

	Prediction		
	Direct	Inverse	Total
Direct	341	27	368
Inverse	88	67	155
Total	429	94	523

Table 5.2: Contingency table of the direction predictors

The fact that the majority-prediction baseline does so well suggests looking more closely at the minority case, inverse scope. The trained models collectively predicted inverse scope with precision $67/94 = 71.2\%$ and recall $67/155 = 43.2\%$ ($F_1 = 53.8\%$).

However, these measures run amiss conceptually by assuming that only inverse scope is of interest. They would be the same if there had been zero correctly predicted direct scopes, or arbitrarily many. But if we want to use scope predictions to build semantic representations, that would have a considerable effect on their reliability. Of course, we can calculate the same measures for direct-scope predictions: precision $341/429 = 79.5\%$ and recall $341/368 = 92.7\%$ ($F_1 = 85.6\%$). But this just leaves out a different class of interest. The two wrongs do not really make a right.

This one-class flaw is among those discussed by Powers (2015). As conceptually similar but more principled replacements for precision, recall, and F_1, Powers argues for markedness, informedness, and Matthews correlation. Markedness amounts to joint precision: Precision for inverse scope plus precision for direct scope, minus one. Informedness amounts to joint recall in the same way. Matthews correlation is their geometric mean. Collectively, the trained systems achieved 50.8% markedness, 35.9% informedness, and 42.7% Matthews correlation.

These (and other) measures can be calculated from the summed contingency table (Table 5.2). Tables for the individual folds are in Appendix B.

5.1.3 Accuracy breakdown by determiner

Although the classifier operated on encodings of nouns (and pronouns), contextual encoding allows for influence from the noun phrase's determiner (if any), and we might well wonder whether this

Determiner	Direct		Inverse		Likelihood
	Correct	Out of	Correct	Out of	
a/an	101	107	15	27	86.6%
cardinal	15	18	3	10	64.3%
demonstrative	10	11	10	19	66.7%
many	23	24	7	14	78.9%
misc.	12	14	4	8	72.7%
most	16	16	0	0	100%
null	257	273	37	110	76.8%
one	10	11	4	5	87.5%
possessive	11	14	1	2	75.0%
some	22	22	0	4	84.6%
the (definite)	49	63	27	45	70.4%
the (generic)	7	7	3	7	71.4%

Table 5.3: Correct predictions given determiner

contributed to accurate prediction. For example, did the models learn not to put 'each' at narrow scope?

I identified the determiners for the two nominals in each data item, and used the predicted and gold scopings to calculate likelihoods conditioned on them. The primary calculation is the probability that an item is predicted correctly given that a certain determiner is present. I wrote a script to tally the successes and failures for each determiner and produce the likelihoods, which I summarize (consolidating extremely small counts) in Table 5.3.

For better perspective, my script also kept more specific tallies and calculated some likelihoods of secondary importance. I kept separate counts for determiners of the first or second nominal in the item, and by their position in the gold scoping. In various combinations with the primary counts, these produced an assortment of other conditions and outcome probabilities.

Even the primary tallies were often small, and the secondary tallies were almost always smaller. For the sake of prudent doubt about estimates of population proportion from small samples, I here report the counts as well as the ratios, and I reviewed the likelihoods as calculated and ranked with various amounts of Laplace smoothing (pseudocounts).

An ad-hoc combination of those smoothed rankings suffices to organize the following discussion, bringing some more noteworthy findings to the front while striking a reasonable balance between the size of the likelihood and the size of the supporting sample. Smoothing also helps to identify a small, relevant subset of the secondary results, since it pulls smaller samples toward 0.5 with greater strength; I mention the secondary findings only when the likelihood exceeds the corresponding primary result under the traditional 'plus four' estimate (adding two successes and two failures; Wilson, 1927).

5.1.3.1 a/an

The count in the table includes 'another', which was predicted correctly 10/11 times (90.9%). Without it, accuracy on 'a/an' falls very slightly to 106/123 (86.2%).

Grouping 'a/an' with other existential determiners (bound and free 'some', 'one') made no particular difference because 'a/an' dominates the combined counts.

Many of the determiners, including some with respectable sample sizes, have a preferred side of the data item; 'some' and 'many' quantify the first noun in the item much more often than the second one, whereas the null determiner of bare mass nouns and the generic use of 'the' strongly favor the second noun. Somewhat unusually, 'a/an' appeared almost equally often in both positions (57 first vs. 75 second), and moreover was predicted correctly at almost identical rates for both.

5.1.3.2 most

The quantifier 'most' was always on the first noun, always at high scope, and always predicted correctly. The sample size of only 16 tempers the accomplishment. So does the fact that the behavior to be learned is just not to depart from the majority prediction.

5.1.3.3 The null determiner

Though without an overt quantifying word, pronouns (when serving as discourse anaphora), proper nouns, bare plurals, bare mass nouns, and even a scattering of bare count nouns are understood with

Determiner	Direct		Inverse		Likelihood
	Correct	Out of	Correct	Out of	
bare singular count noun	11	12	0	7	57.9%
bare mass noun	91	97	23	39	83.8%
bare plural noun	112	115	16	49	78.0%
pronoun	91	95	0	20	79.1%
proper noun	24	27	4	7	82.4%

Table 5.4: Correct predictions given null determiner type

quantificational force and occur in necessary scopal relations.

The understood quantificational force varies. When proper nouns do have a single unique referent, existential and universal force are indistinguishable. Singular pronouns are similar. In Simple English Wikipedia, plural pronouns, bare plurals, and bare mass nouns are often generic as subjects or existential as objects.

Ignoring these distinctions, and collecting all of their outcomes on the basis of their surface similarity, creates a category that appears in both the first and the second positions of data items in very equal balance, is associated with both scopings in fairly equal balance, and is predicted correctly at a rate of 294/383 (76.8%). Some of the individual classes are not much better; bare plurals manage 128/164 (78.0%) and pronouns 91/115 (79.1%). Table 5.4 gives the details. Summing through the table exceeds the total counts given for this category above because data items containing two of these classes will counted twice.

Pronouns are predicted somewhat more accurately when they appear in their typical, first position (80/97, 82.5%), and especially when they are at their typical high scope (80/92, 87.0%). This appears to combine a learned lexical regularity with the direct-scope bias.

Bare plurals have a similar small gain in accuracy from being in second position, and they cannot be said to have a typical position, but they have a clear tendency toward low scope, so direct-scope bias probably accounts for this.

Bare mass nouns do somewhat better to begin with, at 114/136 (83.8%). They have a much

weaker tendency toward second position (63 items vs. 51 for first position, plus 22 items with bare mass nouns in both), and hardly any tendency to a particular scope (59 items low, 51 high, and the 22 doubles), and yet are predicted better in the slightly more common cases: second position 70/81 (86.4%), low scope 67/77 (87.0%). This then may reflect more adequate training for these cases, and not just direct-scope bias.

5.1.3.4 some

'Some', including seven instances of 'someone' and five of 'something', was predicted correctly in 22/26 items (84.6%), including all seven 'someone's but only three 'something's.

All four mispredictions were of direct scope when the correct scope was inverse: Two instances of high-scoped 'something' in second position, and two instances of (the word) 'some' low-scoped in first position. All 22 correct predictions were of direct scope. The correct prediction rate for both first position and high scope therefore rises to 17/19 (89.5%), a respectable number but one based on a small sample, and again a case where fairly consistent training data reinforces the default.

5.1.3.5 one

Though rare relative to 'a/an', 'one' appeared three times more often than all other ordinals put together, and with a strikingly high proportion of accurate predictions: 14/16 (87.5%). For what small samples are worth, 8/8 were correct with 'one' scoped low, and 6/8 with it scoped high. There was one misprediction in each direction.

5.1.3.6 many

'Many' chalked up a less impressive ratio but a fair sample size: 30/38 (78.9%, just slightly better than the all-items average 78.0%). For the 14 items with inverse scope, it split its predictions evenly, but erred only once on the 24 with direct scope.

This quantifier is skewed toward high scope and toward the first position, and when found in these conditions was predicted correctly more often: 24/29 (82.8%)and 23/27 (85.2%). In the

intersection of these conditions, high in direct scope, the ratio was an excellent 20/21 (95.2%). All of the direct scopes in which it was *low* were predicted correctly also (5/5). But in the 14 items whose gold scope was inverse, the predictions were evenly split: 3/6 when low in inverse scope, 4/8 when high.

5.1.3.7 the (generic)

'The', used with generic meaning, was present in 14 items, with 10/14 (71.4%) correct: When it was in second position, six correct predictions of direct scope and three of inverse; and when it was in first position, one further prediction of direct scope. The four mispredictions were all of direct scope when gold was inverse: Two when it was low in first position, two when it was high in second. The three correctly predicted inverse scopes are a good sign and suggest training was effective for something other than reinforcing the direct default, though again the sample size demands caution.

5.1.3.8 the (definite)

Without claiming to have resolved a century-long debate about whether definite descriptions quantify (among many others Russell, 1905; Strawson, 1950; Kripke, 1977; Fodor and Sag, 1982), as a matter of practice, in the Simple English Wikipedia genre we usually find definite 'the' under the scope of a generic assertion, where its asserted meaning can be formalized as a universal quantifier,[2] and this involves it in scopal interactions.

[2]The rest of the meaning of 'the' is presuppositional: 'Fungi get the carbon they need from other organisms' or 'Each bus stopped and the driver got out' presupposes that needed carbon or bus drivers exist, and that the particular portion of carbon or the particular driver can be uniquely identified, relative to a given fungus or bus.

These are part of the facts about fungi or buses. The reader is expected to understand something like, 'Where all fungi need carbon, a generic fungus needs carbon and gets its carbon from other organisms.' We can formalize accordingly: The restrictor set λ_f(fungus f) becomes

$$\lambda_f(\text{fungus } f \wedge \forall_{f'}(\text{fungus } f' \rightarrow \exists_c(\text{carbon } c \ \wedge \ \text{needs } f' \ c)))$$

i.e. 'fungus, where every fungus needs carbon'. The nuclear scope set can fold this down to

$$\lambda_f(\text{fungus } f \wedge \ \exists_c(\text{carbon } c \ \wedge \ \text{needs } f \ c))$$

or 'fungus that needs carbon'.

This device of pushing presuppositions into the outer scope is not limited to definite descriptions. 'Other organisms' presupposes that for a given fungus, it is an organism and there exist organisms that are not it. These propositions can be handled similarly.

	Prediction		
	Direct	Inverse	Total
Direct	49	14	63
Inverse	18	27	45
Total	67	41	108

Table 5.5: Contingency table for items with definite 'the'

This category excludes ordinals and superlatives, which have been counted separately because of the additional quantification embedded in their semantics.[3] It includes 89 items in which one noun had definite 'the', and 19 items in which both did. See Table 5.5 for the outcomes, which are quite distinct from the average behavior (Table 5.2).

Correct predictions in the presence of 'the' were well below average (76/108 or 70.4% vs. 78.0% overall), partly because the frequency with which inverse scope was predicted was more than twice the average. The aforementioned pattern of a generic outscoping definite 'the' usually appears as direct scope, so this high frequency is peculiar, and accuracy is certainly not the only statistical measure that suffered for it.

The higher frequency of inverse predictions reduced precision and especially recall of direct scoping (precision 73.1% vs. 79.5% overall; recall 77.8% vs. 92.7% overall; F_1 = 75.4% vs. 85.6% overall).

Fans of robust simplicity can imagine that wherever the predicate 'fungus' goes, it simply totes along all the extra material about carbon—and about other organisms, and every other background fact about fungi. Cognitive scientists and pragmaticists might instead imagine that when the predicate 'fungus' is in working memory, all generalizations about it are easier to retrieve from semantic memory, and the relevant ones are found and imported to its restrictor set when they match a hypothesis triggered by 'the' or 'other' about what facts have been presupposed.

(Definites are usually treated as presupposing maximal reference, as well as existence and identifiability; 'Fungi get the carbon they need' is taken to mean not just some of the carbon a fungus needs, but all of it; similarly for definite plurals. Maximality is not overtly asserted in the fungi/carbon sentence, but unlike existence, it is at-issue. If fungi got half of their carbon from other organisms and half from the atmosphere, the sentence would be wrong, whereas if fungi did not need carbon, the sentence would be not-even-wrong. And, just as many generics seem to make universal claims but allow for pragmatic exceptions, so too does maximality (Schwarz, 2013). For these reasons I tend to view it as a genuine generic, supplied pragmatically as the tacit quantificational force of the noun phrase, not as a part of the outscoper's restrictor like existence and unique identifiability. But, as with many other questions, we do not claim to have settled the debate. all of this treatment of definites is experimental and provisional, maximality included. 'Simplicity' of the wiki notwithstanding, a large corpus of logical forms touches upon all sorts of open semantic and pragmatic questions, and we adopt provisional answers like these in order to move ahead with it at all.)

[3] We model 'the third' as 'the one such that two members of this set precede it', and superlatives similarly.

Precision for inverse scoping also dropped modestly (down to 65.9% from 71.2% overall), but with a considerable rise in recall (to 60.0% from 43.2%; F_1 = 62.8%, up from 53.8%).

The net effect was a heavy loss on markedness (39.0% vs. 50.8%), a small victory on informedness (37.8% vs. 35.9%)), and a loss on Matthews correlation (38.4% vs. 42.7%).

5.1.3.9 Other definite determiners

The accuracy of definite 'the' is compared to other definite determiners in Table 5.6. Superlatives and ordinals that were removed from the count of 'the' appear here, as does 'the first' (which is arguably both). Apart from 'the first', they are grouped because the individual types are extremely rare.

Five items included an ordinal number other than 'one', and were predicted without error. These were grouped in the expectation that few of them would be attested more than once.

The fairly good performance of possessives likely reflects the fact that the possessor is routinely also the immediate outscoper. Speculatively, the fact that they fall close together in word order and in the syntax tree may help to get them predicted alike with respect to other scope-bearers, but much more data would be needed to investigate this.

If demonstratives are grouped, their accuracy is 20/30 (66.7%). 'Those' was not attested. My only worthy hypothesis about the relatively good performance of 'these' is sample size.

5.1.3.10 Cardinal numbers

This heading collects all cardinal numbers other than 'one', on account of their rarity.

The number two was found in seven data items, once as an Arabic numeral and six times spelled out. One of the spelled-out instances was mispredicted as direct scoping when the gold label was inverse. All other instances were correctly predicted as direct. 'Two or more' was also attested once, mispredicted as inverse scoping when the gold label was direct.

'Three' occurred twice and was predicted correctly (in direct scope) both times.

Determiner	Direct		Inverse		Likelihood
	Correct	Out of	Correct	Out of	
the first	1	1	2	2	100%
ordinal	3	3	2	2	100%
possessive	11	14	1	2	75.0%
superlative	3	3	3	5	75.0%
that	1	2	4	7	55.6%
the (definite)	49	63	27	45	70.4%
these	4	4	4	6	80.0%
this	5	5	2	6	63.6%

Table 5.6: Likelihoods of correct prediction given definite determiner

The number 87.969 (which is the orbital period of Mercury in days) appeared in two data items. Both had inverse scope in the gold labeling; one was predicted correctly.

No other cardinal numbers were repeated, but the synonyms ‘at least one’ and ‘one or more’ occurred in one data item each, which were mispredicted as inverse and direct respectively.

The other cardinals ranged from a low of 1.5 to a high of 200 billion, with a median halfway between 70 and 184, and covered eight direct scopes (seven predicted correctly) and six inverse scopes (two predicted correctly).

5.1.3.11 Miscellaneous determiners

The few remaining determiners are summarized in Table 5.7.

‘All’ occurred in five items, but with only three correct predictions. Other attested universals were ‘both, each, every’ and, as it happens, all four tokens of ‘any’,[4] but combining all of their counts still gives a sample too small for comment. Whether RoBERTa can learn the generalization about ‘each’ remains for practical purposes unanswered.

‘A lot of’ was treated as a multi-word quantifier because the context made it clear that ‘lot’ did not denote an entity.

[4]I.e. negative-polarity ‘any’ was absent.

Determiner	Direct		Inverse	
	Correct	Out of	Correct	Out of
a lot of	0	0	1	1
all	3	5	0	0
(bound) any	2	2	0	1
(free) any	1	1	0	0
both	1	1	0	0
each	3	3	0	0
(bound) every	0	0	0	1
(free) every	1	1	2	2
much	0	0	1	1
no	1	1	0	1
several	0	0	0	1

Table 5.7: Accuracies of sparsely-attested determiners

(a) (b)

Figure 5.1: A misprediction creating a cycle

5.2 Cyclicity

Counting misdirected arcs is a reductionist view on mispredictions. A complementary, holistic view considers patterns in the graph.

The truth-conditionally necessary scopings of the gold data form transitively closed directed acyclic graphs. The test sets contain corresponding undirected graphs, to which the predictor must restore direction arc-by-arc. In other words, the predicted graph can be generated from the gold graph by reversing zero or more arcs.

The most damaging effect this can produce is a directed cycle, as seen in Figure 5.1. A transitive

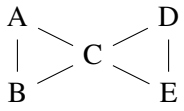

Figure 5.2: Two cycles that count and two compound circuits that do not

triple like 5.1a is a cycle in the undirected graph, which becomes a directed cycle with the misprediction shown in 5.1b. But no lambda expression can be constructed consistent with a cyclic scoping. From the scope-as-continuation perspective of Schuler and Wheeler (2014), we cannot describe the quantification corresponding to any of the three terms until we have described the quantification that points to it. The cycle has to be broken arbitrarily to make a verifiable (true or false) claim.

Producing few or no directed cycles is therefore a requirement for any arc-by-arc scope predictor.[5] Any approach that is cycle-prone needs to be extended to break them (perhaps by examining the weights for each arc prediction, as the greedy approximation algorithm of Manshadi et al. 2013 does), constrained to avoid attempting them (perhaps by only predicting the arcs of the graph's transitive reduction), or abandoned.

In fact, the predictor produced no directed cycles. The magnitude of this accomplishment depends on just how many cycles the predictor *could* have produced.

5.2.1 Count of undirected cycles

A directed cycle can only be predicted where the undirected graph of test items contains a cycle, so a count of undirected cycles helps to contextualize the success of acyclic predictions.

I count cycles in the graph-theoretical sense, so all rotations (and both reflections of an undirected cycle) are equivalent. The definition of cycles also excludes compound circuits, such as ABCEDCA and ABCDECA on Figure 5.2, which make multiple uses of a single vertex.

These two circuits hint at my reason for excluding compounds. Using the point where cycles touch (C) to divide the graph into trails, two distinct compounds are formed by reversing one of the

[5]Disrupting the gold graph's other property, transitive closure, turns out to be unimportant to measure. Discussion is in Appendix C.

trails.[6] As slightly larger numbers of cycles touch each other, the options for reversing and skipping trails increase, and compound circuits proliferate wildly.

However, checking for directed compound circuits amid this proliferation, instead of for directed cycles, does not improve our ability to spot faulty graphs. A directed compound circuit certainly causes the same interpretability problems as a directed cycle, but it also necessarily includes at least two directed cycles, so finding cycles suffices to warn us that the graph is bad.

Moreover, the fraction of compound circuits that are predicted as directed compound circuits[7] is not a better measure of a faulty prediction's faultiness than the fraction of cycles is. If anything, it is worse.

Note that it is impossible to assign directions to the edges of Figure 5.2 such that ABCEDCA and ABCDECA are both directed circuits. A trail they share is reversed from one circuit to the other,[8] so if the directed edges are right for entering C and leaving it again along one of the circuits, they are wrong for the other one. Any way the directions run, at least one of the two gets credit for not being a directed circuit.

In a graph with many compound circuits, there are many pairs of circuits in a similar relationship, and one circuit may participate in several such pairs. This effect drives down the fraction predicted to be directed circuits no matter how badly the directions are predicted. By counting only cycles, I blunt its impact on my measure of quality.

The direction data offered 144 undirected cycles for the predictor to avoid. For additional clarity, I surveyed the documents that produced them and the predictions that were made.

5.2.2 Document sources of undirected cycles

Of the 511 documents that were successfully processed, 210 had one or more truth-functionally necessary outscopings, but only ten had an undirected cycle.

[6] "Reversing" is relative to some other part of the circuit, because of equivalence under reflection. To avoid perpetual double vision, it is convenient to treat undirected cycles/circuits in terms of a canonical forward direction assigned by arbitrary rule. Here, I've made "forward" the direction from A to B.

[7] Or, *mutatis mutandis,* all circuits both simple (i.e. cycle) and complex.

[8] Again, relative to whichever trail is arbitrarily deemed to run "forward" in both.

Cycles per document	Documents
0	200
1	5
3	2
6	1
36	1
91	1

Table 5.8: Distribution of undirected cycles across documents

These documents are distributed fairly evenly across folds: two each in three folds; one each in four folds; none in the two folds that remain. However, the distribution of undirected cycles across documents is notably skewed, as outlined in Table 5.8.

Two unusual sentences were extremely cycle-rich:

(1) Farming is growing crops or keeping animals by people for food and raw materials.

(2) The home page of a web site is the document that a web server sends to another computer's web browser application when it has been contacted without a request for specific information.

Both sentences are noun-heavy by Simple English Wikipedia standards, Example (1) because it is so densely packed (a more typical rate is one noun per three words), and Example (2) because it is so long (around the 95th percentile).

Noun-heavy sentences are usually that way because they include lists, which build scoping graphs that are broad, shallow, and tree-structured. By contrast, these sentences' scope graphs have structures and depth that give their nouns a large number of non-immediate outscopers, which means that taking the transitive closure builds more edges and thus more undirected cycles.

The conjunctions in Example (1) also gave it a scope graph, shown in Figure 5.3,[9] with more reticulation than usual. Only fifteen edges are possible among six vertices; after closure, this sen-

[9]Transitive reduction is shown.

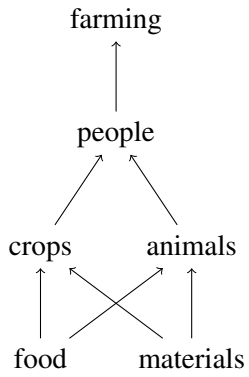

Figure 5.3: Scope DAG for Example (1)

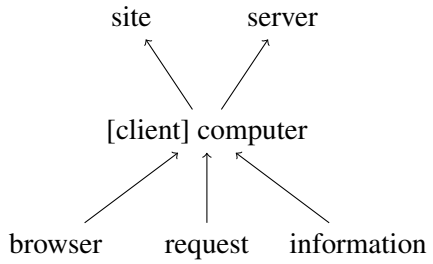

Figure 5.4: Scope DAG for Example (2)

tence's graph had thirteen, more than any other document in the data. This produced the set of 91 undirected cycles.

Example (2) also has a six-noun graph (see Figure 5.4).[10] With less depth and less reticulation, after transitive closure it has eleven edges, second-most of any document in the data set. It produces 36 undirected cycles.

The proliferation of cycles in these two graphs is much like the proliferation of compound circuits discussed in Section 5.2.1, in that the fraction of them predicted as directed cycles is not a very straightforward measure of bad predictions' badness. Fortunately for us, the fraction is zero, which is the most straightforward value it could take.

Zero would be the value if the gold labeling were reproduced exactly by the predictor. But it is

[10] Again in transitive reduction.

also the value under assorted mispredictions. Out of 24 test items related to these two graphs, only 22 were predicted correctly. The two mispredictions just did not happen to be in positions where they could (jointly or singly) form a directed cycle.

The mispredicted items had gold inverse scope and were mispredicted as direct. The 22 items correctly predicted were also direct scope. So in fact, on these items the trained predictor performed identically to the baseline predictor.

This highlights the other factor in cyclic predictions: inverse scope. The baseline cannot predict directed cycles for the simple reason that it never assigns upward scopes that point rightward in the document, only leftward.

5.2.3 Predictions in undirected cycles

To be vulnerable to directed cycles, a predictor needs to predict scopes in both directions. But this is also necessary to beat the baseline on accuracy. To anthropomorphize, a predictor can avoid the challenge of cycles by simply avoiding inverse scope, at a certain cost in accuracy. By predicting inverse scope, it engages this challenge, potentially either well (where inverse scope is correct) or at unnecessary risk (where the correct scope is direct).

Inspecting all the data items that participate in undirected cycles lets us quantify the predictor's tendencies to be bold or foolhardy. One of the documents, "Sausage" in fold 3, had a single cycle of three edges only because of a processing error[11] and I exclude it from further discussion. I also exclude test items from the same documents (and even the same connected components of the graph) that do not themselves form part of any cycle.[12] The remaining items can usefully be divided into two subsets by referring to the gold directed graph.

If reversing a single edge causes a directed cycle, consistency with other predictions weighs heavily in getting it right. These are the edges that are implied by other edges through the transitive property. They are vulnerable to cycles because of this effect when reversed. In our data they are

[11] Scoping was incorrectly assigned among several conjoined examples of types of meat products.
[12] There are nine of these, all predicted with direct scope, eight of them correctly.

	Prediction		
	Direct	Inverse	Total
Direct	19	0	19
Inverse	2	2	4
Total	21	2	23

(a) More cycle-associated

	Prediction		
	Direct	Inverse	Total
Direct	16	0	16
Inverse	11	3	14
Total	27	3	30

(b) In transitive reduction

Table 5.9: Contingency tables for edges included in undirected cycles

almost necessary for cycles also; only three of the undirected cycles do *not* include any of these edges (the 4-edge cycles formed from the six lowest edges in the "Farming" graph, Figure 5.3). There are 23 edges in this more cycle-associated subset.

The other edges are those in the graph's transitive reduction.[13] A single error on one of these edges results in a different acyclic directed graph, but cannot produce a directed cycle.[14] At least two reversals are required. For example, a transitive triple can be turned cyclic either by reversing the redundant edge or by reversing both of the basic ones. Edges belonging to the transitive reduction carry less of the freight of cycle avoidance because of this, and because only the three undirected cycles (previously mentioned) occur solely within this subset. There are 30 such edges.

Counts of predictions, broken out by these subsets, are in Table 5.9. If one considers the six lowest edges in "Farming" cycle-associated and wishes to move them to that subset, all six were predicted as direct scope, correctly in four cases and incorrectly in two.

Mispredictions were present in sufficient numbers to allow for directed cycles either by single reversals of more cycle-associated edges or by multiple reversals among the others. The two single reversals among more cycle-vulnerable edges were salvaged by other mispredictions that restored consistency at a small cost in accuracy; each was a member of a transitive triple including multiple inverse scopings, all of which were predicted as direct. Among edges of the transitive reduction, the mispredictions never fell in dense enough arrangements to set up further transitive cycles.

[13] I.e. the smallest subset of the graph that has the same transitive closure.

[14] Informal proof: Any configuration in which a single edge reversal creates a directed cycle can be derived by starting with a directed cycle and reversing one edge. But after the reversal, it is redundant; its endpoint can be reached from its start by traversing the other part of the cycle. By definition, no such edge is in the transitive reduction.

	Prediction		
	Scope	None	Total
Scope	6	61	67
None	15	2144	2159
Total	21	2205	2226

Table 5.10: Contingency table of the screening predictor

In anthropomorphic terms, the predictor has been cautious, avoiding cyclicity mostly by taking a bias toward direct scope. It has great precision on inverse scope, but sacrifices recall of inverse and precision of direct in order to get there. Broadly speaking, on these and similar statistics its behavior on this portion of the data is consistent with the rest of the dataset.[15] This is to be expected, since the training process did not include loss factors for cyclic predictions (or any other holistic phenomenon).

5.3 Screening subtask

In contrast to the direction subtask, the screening subtask met with little success. The trained model achieved an accuracy of 96.59% in its test set, but the majority-prediction baseline beat it with 96.99%. Table 5.10 counts outcomes by the predicted label (columns) and gold label (rows), where "Scope" indicates a truth-conditionally necessary scopal interaction.

Prediction of interaction had a precision of 0.2857; markedness (combined precision for interaction and non-interaction) was 0.2580. Recall of interaction was 0.0896, and informedness (combined recall) was 0.0826. F_1 was 0.1364 and MCC was 0.1460. For comparison, the majority-prediction baseline scores -0.0301 for markedness, and the remaining statistics are variously zero or undefined.

The screening subtask was not tested on the other folds, because of this poor result and because of other indications that the current data are inadequate to train it. During the hyperparameter grid

[15] The exact percentages are somewhat more extreme, but then, the sample is small.

search, validation accuracies tended to top out early regardless of the values being tried. A logistic regression across all of the validations in the grid search determined that the probability of a new best score fell steeply as training steps increased; at 16,000 minibatches the probability was only $3.3 * 10^{-}3$, and at 25,000 minibatches it sank to $2.2 * 10 - 4$.[16]

Moreover, peak validation scores were always fairly close to baseline performance, with very little effect from hyperparameter tuning. Some did beat the baseline, but the ratio of those that did to those that did not was unconvincing.

These dismal results were achieved in spite of having 37 times more data than the direction subtask and training for 37 times as long. I therefore set the screening task aside.

I interpret these indications to mean that the training process quickly extracted all the information that could be extracted from the available data by an encoder self-trained on RoBERTa's tasks and fine-tuned solely on this one. If so, improving screening performance will require more scope data and/or more data for training other related tasks.

In the meantime, be it noted that writing a lambda expression usually requires adding scopings without truth-conditional effect to the true scopal interactions this subtask is meant to detect. In other words, meaningful practical applications include and accommodate underscreening. Underscreening would increase the opportunities for the direction task to make mutually contradictory predictions, but it has avoided that error very successfully so far.

[16]These probabilities pertain to validations performed at intervals. If I had evaluated after each minibatch, the corresponding probabilities would be vastly lower.

Chapter 6

Remarks and conclusion

This thesis reports the application of the RoBERTa language model to quantifier scope disambiguation, framed as a span pair classification problem with outscoping treated as a semantic dependency between words.

The problem can be expected to fall within the model's abilities on grounds that the model encodes properties of lexis, syntax, and semantics that correlate with human scoping judgements ('scoping factors'). The results of applying similar models to other long-distance, context-sensitive semantic dependency prediction problems are reviewed.

Previously published scope-annotated corpora and scope prediction systems are surveyed. It is found that the systems, or the corpora themselves, either do not cover all of the scoping factors, do not apply them to the full set of quantifiers, or do not represent the full range of subject-matter domains in which humans routinely predict quantifier scope.

The thesis reports development of a new, broad-domain quantifier corpus. Texts are selected for linguistic and domain diversity from a crowdsourced encyclopedia. Training materials, a work process, and the annotator-facing data format were each designed to reduce barriers to entry and safeguard accuracy, with revisions resulting from an inter-annotator agreement study and error analysis. Truth-conditionally meaningful scopal interactions are identified, extracted, and processed into a collection of scope prediction problems.

The thesis discusses appropriate measures of agreement, both between human annotators and between predicted and gold labels, for data having internal structure as a document's scope annotation must. For appropriate calculation of chance-corrected agreement between human annotators,

an inter-annotation distance metric is introduced and justified. For evaluation of automated predictions, where human-like constraints on the structure of a set of predictions are not enforced, results are evaluated both for small-scale accuracy and for compliance with these holistic constraints.

Where scopal interaction is known to exist and only the direction of scope is in question, always predicting direct, or in-situ, scope is a strong baseline. Predictions from the RoBERTa system are shown to be more accurate, to a degree not due to chance, although the system's predictions of inverse scope are still too cautious. The system successfully complies with the holistic constraints, avoiding cyclical outscoping predictions that would block complete semantic interpretation.

6.1 Concerns, next steps, and prospects

The system's narrow margin over the baseline is of concern. So also are the facts that it does not handle the full 3-way prediction problem, in which 'no interaction' is a valid (and the most common) label, or the screening problem of identifying scope interaction without predicting its direction. Next steps to address these, aside from expanding the data (now underway), should include adopting RoBERTa-large in place of RoBERTa-base. With twice as many layers in the encoder, it is likely much better equipped to recognize and extract high-level semantic and pragmatic dependencies. This may make it possible not only to predict scope direction more accurately, but to screen for scope at better than chance levels, and even to re-merge the subproblems and do the full 3-way prediction in a single trained system.

From an engineering perspective, computers to date have little ability to handle the scopings we imply and infer in language about generalizations. Advancing this ability means making them able to engage with more of our linguistic abilities. It is a step toward being able to tell a machine what we know, without having to learn a special formal language for talking to it first. But in the long run, there may be benefits from a scientific perspective too. Working with a system like this may help us to study human cognition of scoping.

Linguists are familiar with formal theories that model human phonological or syntactic computations, but may only loosely approximate our algorithm and may not even attempt to describe our implementation of it *in vivo.* Despite stopping in the upper reaches of the Marr (1982) hierarchy (or rather, continuum; see Poggio 2012), these theories can send us back to human subjects with new predictions to test and new questions to ask, precisely because they stand apart from the linguistic awareness we have when simply using language competence intuitively.

A trained predictor of human scoping judgements is also a theory of their computation, even though a statistical and (technically) neural predictor's theory is a tacit one. The fact that its internals are unlike ours does not invalidate that theory, it just positions it higher on the Marr continuum. Its areas of success or failure may tip us off to previously unrecognized patterns in our scoping judgements. Moreover, since the encode-and-classify architecture is able to share its encoder with multiple other tasks, joint training (particularly adversarial training, designed to burn abilities *out* of the encoder) followed by scoping accuracy evaluation can reveal what kinds of information are most useful for emulating human scoping judgements. This is certainly no proof that human scoping judgements rely on that information, but it is a good place to look for credible hypotheses about the computation that can be tested and walked down the continuum.

In spite of its architectural alienness—twelve attentions and hundreds of simultaneous readers—the encoder is at heart a word-prediction machine, and human language computations apparently have a lot to do with prediction (Schrimpf et al., 2021). So there is a meaningful possibility that what helps a 'Martian' imitate us is, if not what we use, at least connected to it.

6.1.1 Formal theories of scope islands

In 'Rethinking scope islands', Barker (2021) provides an explicit formal description of unavailable inverse-scope readings, which suggests an investigative use for this thesis's corpus and avenues for constructing further scope predictors.

Barker traces the fortunes of the idea that clauses are scope islands, such that a quantifier or other scope-bearer originating within a clause cannot outscope one originating outside it, beginning

with a strained analogy between prohibited inverse scopings and prohibited syntactic movement in the heady 'All grammar is syntax' days of the mid-1970s. He catalogues the strains it suffers (backing up the challenges in the literature with additional naturally occurring examples) and the attempts to salvage it, before breaking apart the island-or-not binary into multiple levels to good descriptive effect. I here summarize.

The early 1980s literature abundantly demonstrates indefinites scoping out of clauses. They were immediately relegated to a separate status, and various analyses were devised in the 1990s and 2000s to excuse them from scope island effects: They are not quantifiers but variables, or alternative sets, or they quantify but can be used in a fashion which delivers truth conditions equivalent to wide scope anyway.

But even for non-indefinite quantifiers, the purported scope islands (clauses generally) are not in fact syntactic islands, and the evidence mounted that actual syntactic islands (e.g. relative clauses) are not scope islands either: In the 1980s literature, Mandarin wh-in-situ scopes out of clauses, and universals scope out of tensed comparative clauses. In the 1990s and 2000s literature, universals also scope out of tensed complement clauses, embedded interrogatives, and relative clauses. Barker backs up these reports with additional naturally-produced examples.

In light of all this, Barker revisits the original evidence for clauses as scope islands, and shows that it does not hold up. Rodman's (1976) sentences fail to prohibit inverse scoping if 'every' is substituted by 'each', and what remains of May's (1977) data has a more coherent semantic description than a syntactic one. Subsequent evidence from cardinal quantifiers also dissolves when their semantics are understood to contain multiple quantifications, and then the indefinite part seemingly remains able to escape to wide scope. And with clause-as-scope-island now thoroughly in question, assorted special mechanisms for producing certain scopes in spite of it have little independent motivation.

Barker's new empirical contribution is that the attested pattern of permitted and prohibited inverse scope can be described with a single well-ordered dimension of strength, on which the scope-taker and the embedding predicate both fall:

- Expressive
- 'Only' focus
- Indefinite
- 'Doubt' complement
- Negative-polarity
- 'Claim' complement
- Universal
- 'Make sure' complement
- Downward monotone

Scope-takers listed higher on this list than some domain are permitted to take scope out of it; those listed lower are not. The quantifier raising allowed by the hierarchy is intended to accommodate everything that can invert to higher scope, obviating a collection of proposed mechanisms that detoured around the former island theory's over-strict constraints to allow some scoping. However, the hierarchy's prohibitions are not meant to similarly obviate all proposed mechanisms that forbid a scoping.

The strength hierarchy can be piggybacked on the function types of logical form; a functional type's strength is fixed in the lexicon, or assigned when it is formed by abstracting over a predication, being in the latter case no stronger than the strongest sub-expression containing the bound variable. Previous formalisms either underdistinguished among scope-takers, or under-constrained their ascent to higher scope, or overconstrained it and required a mechanism for exceptions, or could not handle non-clausal scopes.

6.1.2 'Rethinking scope islands' with the corpus

Barker's natural-language examples considerably strengthen his argument against theories in which clausehood itself (or some subtype of it) creates an island (at least for non-indefinites), and as he puts it, 'to make progress we will need to find a way to collect large data sets of scope judgments'. My corpus could be searched for counterexamples to most of the constraints proposed by his hierarchy—those in which the islands are syntactically defined. His description also makes the focus domain of 'only' an island (for everything except expressives, even indefinites!), which may not always be identifiable in our data.

We do know which words were given typographic emphasis by the writer (bold, italic, etc.), which may designate a focus domain as prosody does in speech, or may perform other functions (again, as prosody does). We can also infer focus domain from information-structural choices such as clefting. But writers may not have been familiar with either the wiki markup for typography or the syntactic constructions (or they may have considered the syntax too advanced for 'simple English'), so our coverage of focus domains is less complete.

Barker has made a point, too, of keeping the description in semantic terms where possible and avoiding overcommitment to particular syntactic instantiations of it. This seems prudent, but it leaves the door open for refined theories with additional non-syntactic island types. Such would likewise be troublesome for our corpus.

6.1.3 Rethinking scope predictors

Using this island theory to support a scope predictor, like any other infusion of human expertise into the system, is all right as far as it goes. When disambiguating scope for practical purposes, any advantage is a good one; if consulting a parse tree can rule out misreadings at less cost than running inference on word embeddings, so much the better. Introducing a new subsystem always has complexity costs, but parsing is getting cheaper all the time, so for the foreseen near future the balance is favorable.

This is assuming the theory is correct. It is undoubtedly less wrong than its predecessor, but I maintain reservations about sentences out of the blue. Judging that 'Someone thought everyone left' must scope in-situ is easy, until I precede it with 'Ella thought Louis left, when he was just out having a smoke. Louis thought Iola and Dave left, when they were in the green room writing. Billie thought Ella and Louis left. In fact—'.

For science, introducing another subsystem is unappealing unless there is some principled reason for it. Citing Steedman (2012), Mike White (personal communication) proposes that this predictor's low performance results from data sparsity, and that scope island restrictions alleviate sparsity (for humans as well as computers) by suppressing a combinatorial explosion of possible analyses. Confirming a sparsity problem with the predictor would then motivate adding such a system.

In the absence of such an indication, an island-based prefilter may still be expedient, just as separating the scope prediction problem into screening and direction was. If a predictor avoids island-violating errors no better than it avoids other errors, we may well ask why. Filtering the data to investigate two classes of errors separately is legitimate and useful. But for a system built on BERT-like representations, the island-compliant dataset does nothing to address the question, 'If embeddings have all this syntax in them, what's stopping the classifier from learning which verbs' complements trap which quantifiers?' The informational capacity of these high-dimensional vectors is vast, so it is much likelier there is something to discover about their processing than that the data itself is just too diverse to handle.

6.2 Data and code availability

The corpus data used in this thesis have been released via the website of the Schuler Computational Cognitive Modeling Lab, and further increments to the corpus will follow in the same location. The task data, as extracted from the corpus, rendered into single scopal interactions, and formatted for use in `jiant`, will be available the same way, together with the scripts for preparing them and the `jiant` task plugin for training and testing the predictor. These materials are not as yet properly

arranged to be downloaded from within `jiant`, but this is a logical next step I hope to pursue, for the sake of engaging a larger population of experimenters with the challenge of scope prediction.

References

Alberti, C., Lee, K., and Collins, M. (2019). A BERT baseline for the Natural Questions. *arXiv preprint arXiv:1901.08634*.

Alshawi, H., editor (1992). *The core language engine*. MIT Press, Cambridge, MA.

AnderBois, S., Brasoveanu, A., and Henderson, R. (2012). The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, volume 16, pages 15–28. MIT Working Papers in Linguistics.

Anderson, C. (2004). *The structure and real-time comprehension of quantifier scope ambiguity*. PhD thesis, Northwestern University Evanston, IL.

Andrew, G. and MacCartney, B. (2004). Statistical resolution of scope ambiguity in natural language. *Unpublished manuscript*.

Artstein, R. and Poesio, M. (2005). Kappa^3 = alpha (or beta). Technical Report CSM-437, University of Essex Department of Computer Science, Colchester, United Kingdom.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barker, C. (2021). Rethinking scope islands. *Linguistic Inquiry*, pages 1–55.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul, Turkey.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239.

Blackburn, P. and Bos, J. (2005). *Representation and inference for natural language: A first course in computational semantics*. CSLI studies in computational linguistics. CSLI Publications.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brandenburg, F. J., Gleißner, A., and Hofmeier, A. (2012). Comparing and aggregating partial orders with kendall tau distances. In *International Workshop on Algorithms and Computation*, pages 88–99. Springer.

Brasoveanu, A. (2010). Structured anaphora to quantifier domains. *Information and Computation*, 208(5):450–473.

Chaves, R. P. (2020). What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics, Vol. 3*, pages 20–30.

Chomsky, N. (1956). Three models for language. *IRE Transactions on Information Theory*, 2:113–124.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cooper, R. (1983). *Quantification and syntactic theory*. D. Reidel, Dordrecht, Holland.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, pages 281–332.

Critchlow, D. E. (2012). *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer Science & Business Media.

Da Costa, J. K. and Chaves, R. P. (2020). Assessing the ability of Transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics, Vol. 3*, pages 189–198.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dima, C. and Hinrichs, E. (2015). Automatic noun compound interpretation using deep neural networks and word embeddings. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 173–183.

Dinesh, N. (2010). *Regulatory conformance checking: Logic and logical form*. PhD thesis, University of Pennsylvania.

Dinesh, N., Joshi, A., and Lee, I. (2011). Computing logical form on regulatory texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1202–1212. Association for Computational Linguistics.

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, 8(11):e81461.

Evang, K. and Bos, J. (2013). Scope disambiguation as a tagging task. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Short Papers*, pages 314–320.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Princeton University.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis*, 1952-59:1–31.

Fodor, J. D. (1982). The mental representation of quantifiers. In Peters, S. and Saarinen, E., editors, *Processes, beliefs, and questions*, pages 129–164. D. Reidel.

Fodor, J. D. and Sag, I. A. (1982). Referential and quantificational indefinites. *Linguistics and philosophy*, 5(3):355–398.

Gao, X., Xiao, B., Tao, D., and Li, X. (2010). A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Gillen, K. (1991). *The comprehension of doubly quantified sentences*. PhD thesis, Durham University.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Grudzińska, J. and Zawadowski, M. (2020). A scope-taking system with dependent types and continuations. In *Logic and Algorithms in Computational Linguistics 2018 (LACompLing2018)*, pages 155–176. Springer.

Hausdorff, F. (2005). *Set Theory*. American Mathematical Society.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Higgins, D. and Sadock, J. M. (2003). A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The Goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Hurum, S. (1988). Handling scope ambiguities in english. In *Proceedings of the second conference on Applied natural language processing*, pages 58–65. Association for Computational Linguistics.

Iyer, S., Dandekar, N., and Csernai, K. (2017). First quora dataset release: Question pairs. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.

Kawakami, K. and Dyer, C. (2015). Learning to represent words in context with multilingual supervision. arXiv:1511.04623.

Kempson, R. M. (1977). *Semantic theory*. Cambridge University Press.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

King, J. C. (2004). Context dependent quantifiers and donkey anaphora. *Canadian Journal of Philosophy*, 34(sup1):97–127.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686.

Koller, A. and Thater, S. (2010). Computing weakest readings. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 30–39. Association for Computational Linguistics.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest studies in philosophy*, 2(1):255–276.

Kuno, S. (1991). Remarks on quantifier scope. In *Current English Linguistics in Japan*, pages 261–288. De Gruyter Mouton.

Kurtzman, H. S. and MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural Questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ladusaw, W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations*. The University of Texas at Austin.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Leslie, S.-J. (2015). Generics oversimplified. *Nous*, 49(1):28–54.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.

Malmi, E., Tatti, N., and Gionis, A. (2015). Beyond rankings: comparing directed acyclic graphs. *Data Mining and Knowledge Discovery*, 29(5):1233–1257.

Manshadi, M. and Allen, J. F. (2011). Unrestricted quantifier scope disambiguation. In *Graph-based Methods for Natural Language Processing*, pages 51–59.

Manshadi, M., Allen, J. F., and Swift, M. (2011). A corpus of scope-disambiguated English text. In *Proceedings of ACL*, pages 141–146.

Manshadi, M., Allen, J. F., and Swift, M. (2012). An annotation scheme for quantifier scope disambiguation. In *Proceedings of LREC*, pages 1546–1553.

Manshadi, M., Gildea, D., and Allen, J. F. (2013). Plurality, negation, and quantification: Towards comprehensive quantifier scope disambiguation. In *Proceedings of ACL*, pages 64–72.

Manshadi, M. H. (2014). *Dealing with quantifier scope ambiguity in natural language understanding*. PhD thesis, University of Rochester.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Martin, P., Appelt, D., and Pereira, F. C. (1983). Transportability and generality in a natural-language interface system. Technical Report 293, SRI INTERNATIONAL ARTIFICIAL INTELLIGENCE CENTER, Menlo Park, California.

Martin, S. (2012). Weak familiarity and anaphoric accessibility in dynamic semantics. In de Groote, P. and Nederhof, M.-J., editors, *Formal Grammar*, pages 287–306, Berlin, Heidelberg. Springer Berlin Heidelberg.

May, R. C. (1977). *The grammar of quantification*. PhD thesis, Massachusetts Institute of Technology.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

McKenna, N. (2019). Learning negation scope semantics with structure. Master's thesis, University of Edinburgh.

McVicar, M., Sach, B., Mesnage, C., Lijffijt, J., Spyropoulou, E., and De Bie, T. (2016). Sumoted: An intuitive edit distance between rooted unordered uniquely-labelled trees. *Pattern Recognition Letters*, 79:52–59.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

Michaelson, E. and Reimer, M. (2019). Reference. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition.

Micham, D. L., Catlin, J., VanDerveer, N. J., and Loveland, K. A. (1980). Lexical and structural cues to quantifier scope relations. *Journal of Psycholinguistic Research*, 9(4):367–377.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 746–751.

Mohammadshahi, A. and Henderson, J. (2019). Graph-to-graph transformer for transition-based dependency parsing. *arXiv preprint arXiv:1911.03561*.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In Hintikka, J., Moravcsik, J., and Suppes, P., editors, *Approaches to Natural Langauge*, pages 221–242. D. Riedel, Dordrecht. Reprinted in R. H. Thomason ed., *Formal Philosophy*, Yale University Press, 1994.

Moran, D. B. (1988). Quantifier scoping in the sri core language engine. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 33–40. Association for Computational Linguistics.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 2125–2140, Mumbai, India.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Parsons, T. (1990). *Events in the Semantics of English*. MIT Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*.

Peters, M., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018b). Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Phang, J., Yeres, P., Swanson, J., Liu, H., Tenney, I. F., Htut, P. M., Vania, C., Wang, A., and Bowman, S. R. (2020). `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, 41(9):1017–1023.

Powers, D. M. W. (2015). What the F-measure doesn't measure: Features, flaws, fallacies and fixes. Technical report, Flinders University.

Pradhan, S. S., Hovy, E. H., Marcus, M. P., Palmer, M., Ramshaw, L. A., and Weischedel, R. M. (2007). OntoNotes: A unified relational semantic representation. In *ICSC*, pages 517–526.

Prange, J., Schneider, N., and Srikumar, V. (2021). Supertagging the long tail with tree-structured decoding of complex categories. *Transactions of the Association for Computational Linguistics*, 9:243–260.

Pütz, T. and Glocker, K. (2019). Tüpa at semeval-2019 task1:(almost) feature-free semantic parsing. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 113–118.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. online PDF.

Rasmussen, N. E. and Schuler, W. (2020). A corpus of encyclopedia articles with logical forms. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1051–1060, Marseille, France. European Language Resources Association.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, pages 8592–8600.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Rodman, R. (1976). Scope phenomena, "movement transformations," and relative clauses. In *Montague grammar*, pages 165–176. Elsevier.

Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Schuler, W. and Wheeler, A. (2014). Cognitive compositional semantics using continuation dependencies. In *Third Joint Conference on Lexical and Computational Semantics (*SEM'14)*.

Schwarz, F. (2013). Maximality and definite plurals-experimental evidence. In *Proceedings of Sinn und Bedeutung*, volume 17, pages 509–526.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sergeeva, E., Zhu, H., Tahmasebi, A., and Szolovits, P. (2019). Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187.

Skjærholt, A. (2014). A chance-corrected measure of inter-annotator agreement for syntax. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics—Long Papers*, volume 1, pages 934–944.

Srinivasan, P. and Yates, A. (2009). Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3, pages 1465–1474. Association for Computational Linguistics.

Steedman, M. (2012). *Taking Scope - The Natural Semantics of Quantifiers*. MIT Press.

Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.

Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Tarski, A. (1933). The concept of truth in the languages of the deductive sciences (polish). *Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych*, 34. translated as 'The concept of truth in formalized languages', in: J. Corcoran (Ed.), Logic, Semantics, Metamathematics: papers from 1923 to 1938, Hackett Publishing Company, Indianapolis, IN, 1983, pp. 152–278.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., and Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*. arXiv:1905.06316.

Tratz, S. and Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687. Association for Computational Linguistics.

Tsiolis, K. C. (2020). Quantifier scope disambiguation. *Unpublished manuscript*.

Tunstall, S. L. (1998). *The interpretation of quantifiers: Semantics & processing*. PhD thesis, University of Massachusetts at Amherst.

van der Sandt, R. A. (1992). Presupposition projection as anaphora resolution. *Journal of semantics*, 9(4):333–377.

VanLehn, K. A. (1978). Determining the scope of English quantifiers. Technical Report AI-TR-98, MIT, Cambridge, Massachusetts.

Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, R. T., Patel, R., Kim, N., Tenney, I., Huang, Y., Yu, K., Jin, S., Chen, B., Van Durme, B., Grave, E., Pavlick, E., and Bowman, S. R. (2019). Can you tell me how to get past Sesame Street?: Sentence-level pretraining beyond language modeling. arXiv:1812.10860. Supersedes Bowman et al. 2018 (same authors), "Looking for ELMo's Friends".

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2020). SuperGLUE: A stickier benchmark for general-purpose language understanding systems.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Warstadt, A. and Bowman, S. R. (2019). Grammatical analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretič, P., and Bowman, S. R. (2019a). Investigating BERT's knowledge of language: Five analysis methods with NPIs.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2019b). BLiMP: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.

Warstadt, A., Singh, A., and Bowman, S. R. (2019c). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Warwick, K. and Shah, H. (2016). Can machines think? a report on Turing test experiments at the Royal Society. *Journal of experimental & Theoretical artificial Intelligence*, 28(6):989–1007.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Woods, W. A., Kaplan, R., and Nash-Webber, B. (1972). The lunar sciences natural language information system: Final report. Technical Report 2378, Bolt, Beranek and Newman, Cambridge, Massachusetts.

Zeng, Z., Tung, A. K., Wang, J., Feng, J., and Zhou, L. (2009). Comparing stars: On approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1):25–36.

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Zhou, Y. and Srikumar, V. (2019). Beyond context: A new perspective for word embeddings. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 22–32.

Appendix A

Language models as multitask learners

Radford et al. (2019) provocatively title their paper, 'Language models are unsupervised multitask learners'. On closer inspection of their results, I have to qualify this claim: Language models are unsupervised multitask learners insofar as the language that is modeled represents the cognitive content of the task, and the training data is sufficient for this cognitive content to shine through the language.

Outside of these conditions, language models are effective scavengers of interesting information that can support task performance, but there are meaningful NLP tasks for which 'Given a question as context, find the linguistically most probable answer' is no more a general solution than 'Pretend to be a 13-year-old Ukrainian boy' (Warwick and Shah, 2016) is a general solution to artificial intelligence.

Radford et al. (2019) attempted zero-shot transfer learning to several downstream tasks with a suite of four language models spanning an order of magnitude in size (from 117M parameters to 1542M, dimensionalities from 768 to 1600). Two of the four were larger than the largest BERT model of Devlin et al. (2018). Given the size advantage, it is unsurprising that the largest of these set a new state of the art for most of the language modeling tasks *strictu senso.* Other downstream tasks were cast in language-modeling terms, but retained significant non-linguistic task requirements, and merely establishing an enormous language model does not seem to have conferred much leverage against these.

In the Children's Book Test (Hill et al., 2015), where the task is selecting correct mid-sentence cloze fillers from a slate of ten options, all four models beat the state-of-the-art (SOTA) accuracy both for common-noun fillers and for named-entity fillers. In this setting, modeling word probability

amounts to modeling the entities in the preceding discourse and their likely activities and roles. Given the models' vast capacity, it is unsurprising that they were able to manage this information, and given how closely the downstream task resembles the pretraining task, it is not surprising that the training transferred well.

In LAMBADA (Paperno et al., 2016), the task is again filling a blank, this time sentence-final and considerably more difficult. Human judges are unable to fill the blank given the sentence alone, but are able to do so given the preceding 50 tokens as context. Modeling the probability of this last word therefore amounts to modeling the topics or themes of a discourse. Here, all four models improved on SOTA perplexity, but only the larger two improved on SOTA accuracy.[1] It is again not surprising either that language modeling was good training, or that a fairly large model was necessary.

The Winograd Schema challenge (Levesque et al., 2012) is a task of context-dependent pronoun resolution. Each item contains a pronoun with two candidate antecedents, a possible single-word substitution that would cause the preferred candidate to change, and a natural-language question asking about the pronoun's reference. This is couched as a language modeling task by inserting the candidate antecedents in place of the pronoun and calculating the probability of the resulting sentence. Here modeling language implicitly models knowledge about the properties of the entities language represents. Once again the larger two models improved on SOTA.

Conversation Question Answering (Reddy et al., 2019) requires disambiguating questions within a dialog and returning the correct answer from a preceding document. Training data include not only an answer but a span of the document that justifies it. This is only incidentally a language generation task, just as giving an answer by moving a robot arm to point at the document would be only incidentally a motor control task. The bulk of the problem lies in reading comprehension and pragmatic reasoning.

Radford et al. presented results only from their largest system, which beat three out of four supervised baselines with an F1 of 55% and came nowhere near (supervised) SOTA. Error analy-

[1] This curious outcome was possible because SOTA for perplexity and accuracy were from two different systems.

sis suggested its reading comprehension and pragmatic reasoning were extremely shallow. This is understandable, since its Internet-derived training data would have included few examples of documents immediately followed by multi-turn question-and-answer dialog, but it does illustrate the limitations of "everything can be a language modeling problem." Everything can be a language modeling problem if there is sufficient linguistic demonstration of the underlying task.

The CNN/Daily Mail summarization tasks (Hermann et al., 2015; Nallapati et al., 2016) illustrate that "sufficient linguistic demonstration" can be a very high bar to clear. Summarization is widespread on the Internet, often conveniently marked "TL;DR" which was a useful cue for the system. Nonetheless even the largest model was competitive only with the trivial technique of randomly selecting sentences from the text, . Its summaries were syntactically acceptable and somewhat on-topic but often factually inaccurate; it did not see through the language of its training data to perceive the underlying task requirement that summaries must be factual.

The language model's data was filtered to exclude non-English web pages, but 10 MB of scattered French passed the filter, which allowed the language model to gloss English sentences in French and to produce baseline-quality English translations from French.[2] This task appears to have been 'language-modely' enough for the system to do well despite the lack of relevant data. I question whether the same would be true for a pair of languages that have not been in close contact with one another, and with the same neighbors, for a thousand years.

Natural Questions (Kwiatkowski et al., 2019) comprise questions originally entered as Google searches, which brought up a suitable Wikipedia page among the top 5 hits. Data includes a short answer and a longer evidence passage from Wikipedia. Here again the language model can only be as good as the reading comprehension model implicit in it. The largest model answered 4.1% of questions correctly, an order of magnitude behind dedicated question answering architectures. So it may not be impossible to train an information extraction model through the veil of a language generation model, but it is certainly inefficient.

To summarize, the more narrowly linguistic a task is, and the more it resembles small-scale text

[2]Using the test set from http://www.statmt.org/wmt14/translation-task.html

generation, the more a language model really is an unsupervised learner for it. Language models capture sufficient information about lexical semantics and semantic prosody to perform well on tasks that require them. But it is not at all the case that a model of talking is necessarily a model of saying the right thing. Even when the task is linguistic, if it is in essence a receptive task, modeling the production of talk about it is a poor proxy for modeling the task itself.

Appendix B

By-fold contingency tables for scope direction predictor

This appendix gives a separate contingency table for the scope direction predictor in each fold of the data.

	Prediction		
	Direct	Inverse	Total
Direct	34	4	38
Inverse	11	18	29
Total	45	22	67

Table B.1: Contingency table of Fold 1 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	20	0	20
Inverse	6	6	12
Total	26	6	32

Table B.2: Contingency table of Fold 2 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	45	3	48
Inverse	14	6	20
Total	59	9	68

Table B.3: Contingency table of Fold 3 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	33	0	33
Inverse	13	1	14
Total	46	1	47

Table B.4: Contingency table of Fold 4 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	36	12	48
Inverse	7	7	14
Total	43	19	62

Table B.5: Contingency table of Fold 5 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	57	2	59
Inverse	14	14	28
Total	71	16	87

Table B.6: Contingency table of Fold 6 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	44	3	47
Inverse	9	3	12
Total	53	6	59

Table B.7: Contingency table of Fold 7 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	42	1	43
Inverse	9	6	15
Total	51	7	58

Table B.8: Contingency table of Fold 8 direction predictor

	Prediction		
	Direct	Inverse	Total
Direct	30	2	32
Inverse	5	6	11
Total	35	8	43

Table B.9: Contingency table of Fold 9 direction predictor

Appendix C

Against scoring predictions for transitive closure

Since the gold standard graph is transitively closed as well as directed acyclic, it was worth considering whether to score the predicted graphs on transitivity. Two kinds of transitivity failures might be scored:

Mispredicted directions can produce paths with no transitive "shortcut," as in Figure C.1, so that the predicted graph is not transitively closed. But this does not leave the graph uninterpretable, as predicting a cycle does, and there is no special reason these are worse disruptions of the semantics than mispredictions that do preserve transitive closure, such as that seen in Figure C.2.

Mispredictions can also destroy transitive triples that are present in the gold graph. But these are also caught by accuracy. Their only further significance is if they create a cycle, which we are already measuring in its own right, instead of just creating a different triple.

In fact, both of these are strict subsets of accuracy violations, and their intersection is a strict

Figure C.1: A misprediction destroying transitive closure

Figure C.2: A misprediction preserving transitive closure

superset of cyclicity violations,[1] so if they have no particular significance of their own for the predicted semantics, we might as well leave them alone.

[1]Transitively closed directed cycles exist, but cannot be reached by arc reversal. Transitively closing a directed cycle requires arcs in both directions between any two nodes in the cycle, and the gold graph has at most one arc between any two nodes.