Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1051–1060
Marseille, 11–16 May 2020

A Corpus of Encyclopedia Articles with Logical Forms

Nathan Ellis Rasmussen, William Schuler
Department of Linguistics
The Ohio State University
{rasmussen.63,schuler}@ling.osu.edu

Abstract

People can extract precise, complex logical meanings from text in documents such as tax forms and game rules, but language processing systems lack adequate training and evaluation resources to do these kinds of tasks reliably. This paper describes a corpus of annotated typed lambda calculus translations for approximately 2,000 sentences in Simple English Wikipedia, which is assumed to constitute a broad-coverage domain for precise, complex descriptions. The corpus described in this paper contains a large number of quantifiers and interesting scoping configurations, and is presented specifically as a resource for quantifier scope disambiguation systems, but also more generally as an object of linguistic study.

Keywords: quantifier scope, coreference, logical form, explanatory text, Wikipedia

1. Introduction

People can extract precise, complex logical meanings from explanatory text in documents such as tax forms and game rules, but language processing systems lack adequate training and evaluation resources to do these kinds of tasks reliably. This paper describes a corpus of annotated typed lambda calculus translations (Church, 1940) for approximately 2,000 sentences in Simple English Wikipedia, which is assumed to constitute a broad-coverage domain for precise, complex descriptions. These typed lambda calculus expressions are intended to serve as a theory-neutral formal semantic representation, to which other representations (Kamp, 1981; Heim, 1982) can be translated. Moreover, output from systems trained on typed lambda calculus representations like these may be run with only minor pre-processing as programs in a functional programming language such as Haskell or Ocaml as a way to evaluate input statements as goals in some world model. The corpus described in this paper contains a large number of quantifiers and interesting scoping configurations, and is presented specifically as a resource for quantifier scope disambiguation systems, and also more generally as an object of linguistic study.

2. Related Work

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, as yet, been reduced to a general-purpose algorithm.[1] To model it statistically requires training data that incorporate the relevant psycholinguistic cues: text coherence (Dwivedi, 2013), linear order of scope-bearers, syntactic structure, choice of quantifiers (as *each* vs. *every*), and presumptions of background knowledge (AnderBois et al., 2012) similar to those found in an explanation.

VanLehn (1978) studied syntactic influences on quantifier scoping to improve scope disambiguation in the Lunar Sciences Natural Language Information System (Woods et al., 1972). He reports 'inconclusive' and 'remote' prospects for improving the system. We know of no active work on the system since, and the data may no longer be extant.

In the WSJ Penn Treebank (41,191 sentences total), Higgins and Sadock (2003) found 893 sentences with two quantified expressions other than the determiners *a, an, the;* the news genre is quantifier-poor (Biber et al., 1999, pp. 277–78). Only 348 (39%) had scopal interaction able to affect truth conditions. A second coder agreed with the reference coding on 76% of sentences, at a Cohen's κ of 0.52.

Andrew and MacCartney (2004) mined 305 two-quantifier sentences from logic problems of the Law School Admissions Test (LSAT), either by extracting them directly or by editing them down from more complex sentences. 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, possibly because the genre discourages ambiguity. It certainly minimizes the contributions of world knowledge and discourse context.

Srinivasan and Yates (2009) labeled 92 semi-synthetic quantifier scope disambiguation problems for an experiment in automatically extracting world knowledge from unlabeled text. Although they began with predicates and their arguments found in the Web1Tgram corpus, they imposed the interacting quantifiers *a* and *every,* and formulated the problems in an MRS-like (Copestake et al., 2005) logical form rather than in natural language. This allowed them to demonstrate how world knowledge aids broad-domain scoping, but precludes learning anything about the effects of information structure from the data.

Dinesh et al. (2011) annotated the scoping of quantifiers and other operators in 195 sentences of FDA regulations (average 30 words per sentence, quantifier count variable), a genre specially enriched in scope-taking modal auxiliaries. They predict each scopal operator's most likely outscoper with a maximum-entropy classifier (Dinesh, 2010). The background knowledge presumed is deep but narrow.

[1] Highly successful algorithms are available for certain special cases (Evang and Bos, 2013; Schuler and Wheeler, 2014).

	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 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.

There may be advantages for machine learning from regulations' position where exact legal reasoning meets 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. But regulations must grapple with the complexity inherent in their subject matter, potentially providing better training in the use of background knowledge than logic puzzles, which can be composed in any domain that minimizes ambiguity.

AnderBois et al. (2012) improved data quality in LSAT puzzles by having multiple annotators code each item. They report 358 non-cumulative two-quantifier sentences with at least one quantifier as subject or direct object (the other quantifier may perform some other function). They imply the existence of other annotated data in the corpus, though without mentioning its quantity. The genre continues to limit the usability of the texts as training data for a general-purpose system.

QuanText, by Manshadi et al. (2013), is to our knowledge the most thoroughly developed corpus of scope annotations. It consists of 500 imperative sentences similar to Example (1), giving instructions for manipulating text files.

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

Sentences were derived from tutorials, help documents, a survey of computer users, and crowdsourced descriptions of example data manipulations (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 quantifier scope over such complex materials (Manshadi and Allen, 2011). In addition to scoping itself, QuanText annotates related phenomena such as collective and distributive readings, partitives, and type/token distinctions (Manshadi et al., 2012).

QuanText sentences routinely contain three or more scope-bearers. 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. But having more than two scope-bearers required QuanText to adopt a more complex annotation scheme than previous projects, and this incurred some problems in the methodology for comparing annotations with one another or with machine predictions (see Section 6.1.).

The principal objects of the domain—characters, words, lines, and files—are overwhelmingly in part–whole relationships, so that a very simple heuristic scoping (Schuler and Wheeler, 2014) rivals both the predictions of a sophisticated machine learning system (Manshadi et al., 2013) and QuanText's inter-annotator agreement (Manshadi et al., 2012). Though this confirms the value of world knowledge for scope prediction, again it limits chances to generalize from the annotated data. Furthermore, the QuanText sentences have been edited to be understandable out of the blue, which prevents any investigation of text coherence or other discourse influences on interpretation.

Evang and Bos (2013) 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, and annotated the scoping between the modificand and the PP object. They acknowledge that the narrowly selected syntax and the purely binary annotation limit what can be learned and even what can be annotated. Furthermore, finding only 456 examples in a million-word corpus suggests that the GMB's genres are poorly suited for a scope corpus.

The present work thus represents an improvement on previous scope corpora in its size, its quantifier-rich genre, its broad subject matter, and in the full spectrum of scoping cues it retains as a natural, connected text.

3. Background: Typed Lambda Calculus

The logical form corpus described in this paper uses typed lambda calculus expressions (Church, 1940) as a theory-neutral representation to which other representations can be translated. Morerover, output from systems trained on typed lambda calculus representations like these may be run with only minor pre-processing as programs in a functional programming language such as Haskell or Ocaml as a way to evaluate input statements as goals in some world model. Types for expressions are drawn from:

- **entities** $\mathtt{e}$
- **truth values** $\mathtt{t}$
- **functions** $\alpha \rightarrow \beta$ from input of type α to output of type β

Expressions themselves are composed of:

- **constants** κ in some domain of expressions, here notated in $\mathsf{sans\text{-}serif}$ font;
- **variables** χ over some domain of expressions, here notated in *italics*;
- **abstractions** $(\lambda_\chi\ \varphi)$ of type $\alpha \rightarrow \beta$ of an expression as a function from some variable χ of type α to some expression φ of type β which may contain that variable; and
- **applications** $(\varphi\ \psi)$ of type β of a function expression φ of type $\alpha \rightarrow \beta$ to an argument expression ψ of type α.

4. An Ontology for Encyclopedia Articles

The corpus described in this paper contains typed lambda calculus translations of articles in Simple English Wikipedia. This domain was chosen because of its complex and generally transparent meaning, and because its Creative Commons license facilitates distribution of the corpus. The Simple English edition was adopted because it uses a smaller vocabulary and grammar, which may simplify annotation and yield fewer sparse data effects when used as a training resource.

Entities in this annotation are understood to subsume:

- **ordinary count entities**, like people or buildings, which can be counted;[2]
- **minimal parts** of substances or measures, which can be quantified by ratios;[3]
- **eventualities** (Davidson, 1967; Parsons, 1990), which are minimal regions of space and time defined by a predicate and a set of participants; and
- **numbers**, which provide thresholds for quantifiers, but can also be constrained and quantified over.

Basic function constants in this annotation are restricted to connectives, operators, predicates and the cardinality function.

- **Connectives** (e.g. conjunctions, denoted by '$\wedge$' or comma) in this annotation are functions of type $\mathtt{t} \rightarrow \mathtt{t} \rightarrow \mathtt{t}$, which map a pair of input truth values to an output truth value.
- **Operators** in this annotation are functions of type $\mathtt{e} \rightarrow \mathtt{e} \rightarrow \mathtt{e}$, which map a pair of input entities (such as numbers) to an output entity (number), which allows an unbounded set of numbers to be represented with a bounded set of digit and operator constants.
- **Predicates** in this annotation are functions of type $\mathtt{e} \rightarrow \mathtt{t}$ or $\mathtt{e} \rightarrow \mathtt{e} \rightarrow \mathtt{t}$ or $\mathtt{e} \rightarrow \mathtt{e} \rightarrow \mathtt{e} \rightarrow \mathtt{t}$, which map one or more entities to truth values.
- **Cardinality** $(|S|)$ is a function of type $(\mathtt{e} \rightarrow \mathtt{t}) \rightarrow \mathtt{e}$ which maps a set of entities S to a number (the cardinality of the set).

4.1. Numerical Quantifiers

4.1.1. Generalized Quantifiers

These basic function constants can derive a set of generalized quantifier functions (Barwise and Cooper, 1981) of type $(\mathtt{e} \rightarrow \mathtt{t}) \rightarrow (\mathtt{e} \rightarrow \mathtt{t}) \rightarrow \mathtt{t}$. These functions take a 'restrictor' set (usually described by a common noun occurring as a syntactic complement of the quantifier) and a 'nuclear scope' set (usually described by a verb phrase or other predicate occurring as a sibling of the quantifier phrase) and return a truth value based on the cardinality of the intersection of these sets and (optionally) its relationship to the cardinality of the restrictor set:[4]

$$(\mathsf{some}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x| > 0) \quad (1a)$$
$$(\mathsf{none}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x| = 0) \quad (1b)$$
$$(\mathsf{two}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x| = 2) \quad (1c)$$
$$(\mathsf{all}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x|\ /\ |R| = 1.0) \quad (1d)$$
$$(\mathsf{most}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x|\ /\ |R| > 0.5) \quad (1e)$$
$$(\mathsf{half}\ R\ S) \Leftrightarrow (|\lambda_x\ R\,x,\ S\ x|\ /\ |R| = 0.5) \quad (1f)$$

This produces relatively simple lambda calculus expressions:[5]

(2) Most libraries are public.

$$\mathsf{most}\ (\lambda_x\ \mathsf{prop\ library}\ x)$$
$$(\lambda_x\ \mathsf{prop\ public}\ x)$$

[2] A word is also an entity when mentioned, distinct from the entity or entities described by the word when used. Types and tokens are treated as distinct entities when possible, following Manshadi et al. (2012), but Wikipedia writers seldom observe the distinction.

[3] Link (1983) and others allow non-minimal amounts of substances to be considered entities, but the use of non-minimal substance entities precludes a uniform treatment of common propositional quantifiers like 'half' or '71%' across count and mass nouns.

[4] Here, the ratio of cardinalities of two infinite sets is defined to be the expected ratio of cardinalities of those sets intersected with a random sample of entities D_K, as sample size tends to infinity:

$$|S|\ /\ |R| \overset{\text{def}}{=} \lim_{K\rightarrow\infty} \mathrm{E}_{D_K \sim \pi}\ |\lambda x\ D_K\,x,\ S\ x|\ /\ |\lambda x\ D_K\,x,\ R\,x|$$

[5] Additionally, expressions will use the following functions to provide low scope bindings for variables over eventualities for properties (prop) and relations (reln):

$$(\mathsf{prop}\ f\ x) \Leftrightarrow (\mathsf{some}\ (\lambda_e\ \mathsf{eventuality}\ e)\ (\lambda_e\ f\ e\ x))$$
$$(\mathsf{reln}\ f\ x\ y) \Leftrightarrow (\mathsf{some}\ (\lambda_e\ \mathsf{eventuality}\ e)\ (\lambda_e\ f\ e\ x\ y))$$

These functions will be further generalized to model cardinal quantifiers (Equations 1a–1c, above), and propositional quantifiers (Equations 1d–1f, above).

4.1.2. Cardinal Quantifiers

The annotations described in this paper use a similar representation for cardinal numbers of type $\mathtt{e} \rightarrow (\mathtt{e} \rightarrow \mathtt{t}) \rightarrow (\mathtt{e} \rightarrow \mathtt{t}) \rightarrow \mathtt{t}$, further generalized to include the number itself as an entity argument which can be constrained by other parts of the sentence:[6]

$$(\mathsf{count}_{=}\ n\ R\ S) \ \Leftrightarrow\ (|\lambda_x\, R\,x,\ S\,x| = n) \tag{2}$$

This formulation of numerical quantifiers allows a straightforward analysis of cardinals:

(3) Some legions have 5,500 men.

$$\begin{array}{l}
\mathsf{some}\ (\lambda_x\ \mathsf{prop\ legion}\ x) \\
\qquad (\lambda_x\ \mathsf{count}_{=}\ \mathsf{5500}\ (\lambda_y\ \mathsf{prop\ man}\ y) \\
\qquad\qquad\qquad\qquad (\lambda_y\ \mathsf{reln\ have}\ x\ y))
\end{array}$$

The numbers themselves can be defined in terms of a finite number of constants for digits and addition and multiplication operators, if it is desirable to avoid sparse or unknown constants in model training.

4.1.3. Propositional Quantifiers

The generalized quantifier is similarly extended to define propositional quantifiers with numerical arguments for ratios:[7]

$$(\mathsf{ratio}_{=}\ n\ R\ S) \ \Leftrightarrow\ (|\lambda_x\, R\,x,\ S\,x|\ /\ |R| = n) \tag{3}$$

For example:

(4) Water covers 71% of the Earth.

$$\begin{array}{l}
\mathsf{ratio}_{=}\ \mathsf{.71}\ (\lambda_x\ \mathsf{prop\ part\text{-}of\text{-}earth}\ x) \\
\qquad (\lambda_x\ \mathsf{some}\ (\lambda_y\ \mathsf{prop\ part\text{-}of\text{-}water}\ y) \\
\qquad\qquad\qquad (\lambda_y\ \mathsf{reln\ cover}\ x\ y))
\end{array}$$

Note that Earth is treated as a continuous substance in this sentence, and the scoping of the subject and object is inverted: for 71% of earth particles, at least one water particle covers it.

4.1.4. Non-conservative Quantifiers

Quantifiers are usually assumed to be *conservative*, i.e. defined over the intersection of the restrictor and nuclear scope sets. Non-conservative quantifiers, marked as prime (′) in this annotation, relax this constraint:[8]

$$(\mathsf{ratio}'_{=}\ n\ R\ S) \ \Leftrightarrow\ (|S|\ /\ |R| = n) \tag{4}$$

This relaxation is necessary when cardinalities of disjoint sets or sizes of different objects are compared. For example:

(5) There are 312 times as many arthropods as mammals.

$$\begin{array}{l}
\mathsf{ratio}'_{=}\ \mathsf{312.0}\ (\lambda_x\ \mathsf{prop\ mammal}\ x) \\
\qquad\qquad\qquad (\lambda_x\ \mathsf{prop\ arthropod}\ x)
\end{array}$$

4.1.5. Measure Phrases

Measure phrases can be modeled using non-conservative quantifiers, as ratios of the measure of some object to the measure of some unit, calculated over quantities of measure parts. For example:

(6) The Matterhorn is 4.5 kilometers tall.

$$\begin{array}{l}
\mathsf{some}\ (\lambda_x\ \mathsf{prop\ matterhorn}\ x) \\
\quad (\lambda_x\ \mathsf{some}\ (\lambda_y\ \mathsf{prop\ kilometer}\ y) \\
\qquad\qquad (\lambda_y\ \mathsf{ratio}'_{=}\ \mathsf{4.5}\ (\lambda_z\ \mathsf{reln\ part\text{-}of\text{-}length}\ z\ y) \\
\qquad\qquad\qquad\qquad (\lambda_z\ \mathsf{reln\ part\text{-}of\text{-}height}\ z\ x)))
\end{array}$$

These measure parts are purely mathematical objects, and are distinct from minimal parts of objects themselves. This analysis generalizes to ad-hoc units as well:

(7) Jupiter is 11 Earths wide.

$$\begin{array}{l}
\mathsf{some}\ (\lambda_x\ \mathsf{prop\ jupiter}\ x) \\
\quad (\lambda_x\ \mathsf{some}\ (\lambda_y\ \mathsf{prop\ earth}\ y) \\
\qquad\qquad (\lambda_y\ \mathsf{ratio}'_{=}\ \mathsf{11.0}\ (\lambda_z\ \mathsf{reln\ part\text{-}of\text{-}width}\ z\ y) \\
\qquad\qquad\qquad\qquad (\lambda_z\ \mathsf{reln\ part\text{-}of\text{-}width}\ z\ x)))
\end{array}$$

4.1.6. Explicit Constraints on Numbers

The explicit argument over numbers in the above functions allows quantifiers to have numerical thresholds that may themselves be quantified over and constrained: e.g.

(8) The number of bytes in each memory is a whole power of two.

$$\begin{array}{l}
\mathsf{all}\ (\lambda_x\ \mathsf{prop\ memory}\ x) \\
\quad (\lambda_x\ \mathsf{some}\ (\lambda_n\ \mathsf{prop\ number}\ n, \\
\qquad\qquad\qquad \mathsf{count}_{=}\ n\ (\lambda_y\ \mathsf{prop\ byte}\ y) \\
\qquad\qquad\qquad\qquad (\lambda_y\ \mathsf{reln\ contain}\ x\ y)) \\
\qquad\qquad (\lambda_n\ \mathsf{some}\ (\lambda_m\ \mathsf{prop\ number}\ m, \\
\qquad\qquad\qquad\qquad \mathsf{prop\ whole}\ m) \\
\qquad\qquad\qquad (\lambda_m\ \mathsf{reln\ equal}\ n\ (\mathsf{power\ 2}\ m))))
\end{array}$$

The number could be described in the generalized quantifier function, but the realization of the number as a separate, explicitly quantified variable allows a uniform translation of numbers and non-numbers from noun phrase descriptions. Similar analyses account for 'a different number,' or 'an even number' or quantifiers involving mathematical operations, e.g. 'July has one more day than June,' or comparatives across kinds, e.g. 'a number of credits equal to the number of hours worked.'

4.2. Generics

Many indefinite and bare plural noun phrases, typically high-scoping subjects describing the topic of an article, seem to have a **generic** force similar to a universal quantifier:

[6] Similar definitions exist for ≤, <, >, ≥.

[7] Again, similar definitions exist for ≤, <, >, ≥.

[8] Again, similar definitions exist for ≤, <, >, ≥.

(9) A king is a man who rules a country.

(10) Experiments are tests.

Following Leslie (2015), these annotations do not treat generics as descriptions of types or kinds, but rather as quantifiers over individuals with an underspecified threshold that is dependent on discourse factors:

$$\begin{array}{l}\text{gen } (\lambda_x \text{ prop experiment } x)\\ \quad (\lambda_x \text{ prop test } x)\end{array}$$

Usually this threshold is close to 1.0, but not always:

(11) Mosquitoes carry malaria.

Example (11) does not require that all mosquitoes carry malaria, nor even most mosquitoes.

4.3. Comparatives

Ratios can also generalize to non-numeric comparative quantifiers:

(12) Islands are smaller than continents.

$$\begin{array}{l}\text{gen } (\lambda_x \text{ prop island } x)\\ \quad (\lambda_x \text{ gen } (\lambda_y \text{ prop continent } y)\\ \qquad\qquad (\lambda_y \text{ ratio}'_< \ 1.0\ (\lambda_z \text{ reln part-of-area } z\ x)\\ \qquad\qquad\qquad\qquad (\lambda_z \text{ reln part-of-area } z\ y)))\end{array}$$

4.4. Ranking Adjectives

Several kinds of quantifiers involve implicit quantification over a set of entities to establish relative rankings.

4.4.1. Ordinal Numbers

Simple ordinals define a ranking based on precedence:

(13) April is the fourth month in the year.

$$\begin{array}{l}\text{gen } (\lambda_z \text{ prop year } z)\\ \quad (\lambda_z \text{ some } (\lambda_x \text{ prop april } x)\\ \qquad\qquad (\lambda_x \text{ prop month } x,\\ \qquad\qquad\quad \text{reln in } x\ z,\\ \qquad\qquad\quad \text{count}_= \ 3\ (\lambda_y \text{ prop month } y,\\ \qquad\qquad\qquad\qquad \text{reln in } y\ z)\\ \qquad\qquad\qquad\quad (\lambda_y \text{ reln precede } y\ x)))\end{array}$$

4.4.2. Superlatives

Superlatives are a special case of ordinals identifying the entity that occurs first in some ranking:

(14) Canada is the country with the most lakes.

$$\begin{array}{l}\text{some } (\lambda_x \text{ prop canada } x)\\ \quad (\lambda_x \text{ prop country } x,\\ \qquad \text{count}_= \ 0\ (\lambda_y \text{ prop country } y)\\ \qquad\qquad (\lambda_y \text{ ratio}'_> \ 1.0\ (\lambda_z \text{ prop lake } z,\\ \qquad\qquad\qquad\qquad \text{reln have } y\ z)\\ \qquad\qquad\qquad (\lambda_z \text{ prop lake } x\ z,\\ \qquad\qquad\qquad\qquad \text{reln have } x\ z)))\end{array}$$

This analysis generalizes to continuous quantification in comparatives:

(15) Russia is the largest country.

$$\begin{array}{l}\text{some } (\lambda_x \text{ prop russia } x)\\ \quad (\lambda_x \text{ prop country } x,\\ \qquad \text{count}_= \ 0\ (\lambda_y \text{ prop country } y)\\ \qquad\qquad (\lambda_y \text{ ratio}'_> \ 1.0\ (\lambda_z \text{ reln part-of-area } z\ y)\\ \qquad\qquad\qquad\qquad (\lambda_z \text{ reln part-of-area } z\ x)))\end{array}$$

This analysis also generalizes to mixtures of ordinals and superlatives:

(16) China is the third largest country.

$$\begin{array}{l}\text{some } (\lambda_x \text{ prop china } x)\\ \quad (\lambda_x \text{ prop country } x,\\ \qquad \text{count}_= \ 2\ (\lambda_y \text{ prop country } y)\\ \qquad\qquad (\lambda_y \text{ ratio}'_> \ 1.0\ (\lambda_z \text{ reln part-of-area } z\ y)\\ \qquad\qquad\qquad\qquad (\lambda_z \text{ reln part-of-area } z\ x)))\end{array}$$

which change the count of how many entities in the restrictor set (countries, in the above example) are claimed to surpass the identified entity.

4.5. Discourse Anaphora

Many sentences contain anaphora whose antecedents are quantified noun phrases in other sentences or clauses. Following King (2004), these discourse anaphora are modeled as a type of quantifier which treats the intersection of the restrictor and nuclear scope of its antecedent quantifier as a restrictor set. For example:

(17) The sun has eight planets. They have circular orbits.

$$\begin{array}{l}\text{count}_= \ 8\ (\lambda_x \text{ prop planet } x)\\ \qquad (\lambda_x \text{ some } (\lambda_y \text{ prop sun } y)\\ \qquad\qquad (\lambda_y \text{ reln have } y\ x)),\\ \text{all } (\lambda_x \text{ prop planet } x,\\ \qquad \text{some } (\lambda_y \text{ prop sun } y)\\ \qquad\quad (\lambda_y \text{ reln have } y\ x))\\ \quad (\lambda_x \text{ some } (\lambda_z \text{ prop orbit } z,\\ \qquad\qquad \text{prop circular } z)\\ \qquad\quad (\lambda_z \text{ reln have } x\ z))\end{array}$$

Importantly, this is not equivalent to:

(18) The sun has eight planets that have circular orbits.

because the former is intuitively not satisfied if there are nine planets, one of which has an non-circular orbit, but the latter is.

4.6. Groups

In order to avoid duplication of predicates at the group and individual level, group predicates like ‘between’ or ‘surround’ are given an equivalent distributed analysis in which a single eventuality is shared by multiple entity-level predicates:

(19) April is between March and May.

$$\begin{array}{l}
\textsf{some}\ (\lambda_x\ \textsf{prop april}\ x) \\
\qquad (\lambda_x\ \textsf{some}\ (\lambda_e\ \textsf{eventuality}\ e) \\
\qquad\qquad (\lambda_e\ \textsf{some}\ (\lambda_y\ \textsf{prop march}\ y) \\
\qquad\qquad\qquad (\lambda_y\ \textsf{distrib-between}\ e\ x\ y), \\
\qquad\qquad\quad \textsf{some}\ (\lambda_y\ \textsf{prop may}\ y) \\
\qquad\qquad\qquad (\lambda_y\ \textsf{distrib-between}\ e\ x\ y)))
\end{array}$$

The single eventuality thus ensures the *between* relation applies to all elements of each participant group with a common time and location.

4.7. Modal operators

Modal operators (can, may, must) and temporal operators (always, sometimes, usually, never) in encyclopedia articles are primarily epistemic, describing probabilities of events. Since statements in this corpus are general and do not assume a specific time or world state, these operators are modeled as quantifiers over events:

(20) Cheetahs usually run fast.

$$\begin{array}{l}
\textsf{gen}\ (\lambda_x\ \textsf{prop cheetah}\ x) \\
\qquad (\lambda_x\ \textsf{ratio}_{>}\ \textsf{.5}\ (\lambda_e\ \textsf{run}\ e\ y) \\
\qquad\qquad\qquad (\lambda_e\ \textsf{prop fast}\ e))
\end{array}$$

Quantifiers over events are implicitly over not only past events but also counterfactual past and possible future events. Quantifiers over past events are then explicitly constrained to the past.

4.8. Reciprocal Pronouns

Reciprocal pronouns like 'each other' are implemented within quantifiers over each member of a set as a secondary quantifier over that set excluding that member:

(21) People see each other.

$$\begin{array}{l}
\textsf{gen}\ (\lambda_x\ \textsf{prop person}\ x) \\
\qquad (\lambda_x\ \textsf{gen}\ (\lambda_y\ \textsf{prop person}\ y, \\
\qquad\qquad\qquad \textsf{reln different}\ x\ y) \\
\qquad\qquad (\lambda_y\ \textsf{reln see}\ x\ y))
\end{array}$$

4.9. Negation

Negation is implemented as a quantifier over a variable not appearing in the restrictor or nuclear scope:

(22) Amphibians are not fish.

$$\begin{array}{l}
\textsf{gen}\ (\lambda_x\ \textsf{prop amphibian}\ x) \\
\qquad (\lambda_x\ \textsf{count}_{=}\ \textsf{0}\ (\lambda_y\ \textsf{true}) \\
\qquad\qquad\qquad (\lambda_y\ \textsf{prop fish}\ x))
\end{array}$$

This allows annotated scope associations to include and interact with negation.

5. Annotation Procedure

The annotated data in this corpus is drawn from encyclopedia articles in a 2014 dump of Simple English Wikipedia. The selected articles are those whose title appears most frequently in the full text of the dump, plus a random sample from among the first 450 articles created. In order to ensure a broad domain, the corpus includes only the first three to six sentences of each article.

5.1. Hand-corrected Automatic Syntactic Annotation

Prior to semantic annotation, the corpus is 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. This markup distinguishes composition operations for arguments, modifiers and various non-local constructions such as filler-gap constructions, each of which selectively constrains restrictors or nuclear scopes of quantifiers, depending on the operation. In general, meanings of modifier predicates are applied to restrictors of quantifiers associated with modificands, and meanings of non-modifier predicates are applied to 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 quantified noun phrase, and for any verb, adjective or prepositional phrases that require quantifiers over eventualities. After it is parsed, the corpus is hand corrected to fix automatic attachment errors and to ensure that valid elementary predications can be obtained.

5.2. Hand-specified Semantic Associations

Annotators then mark preterminal nodes of head words of noun phrases, modal auxiliaries and negation modifiers in these parsed trees to specify anaphor antecedents and scope parents of quantifiers. Intuitively, anaphor antecedents are pointers to the most recent word that describes the same entity or entities described by the marked word. Scope parents, marked on one (low) noun to point to another (high) noun, specify that there is a set of entities described by the low noun for each entity described by the higher noun. The annotator's task then amounts to drawing arrows, after which anaphor antecedents and scope parents are automatically validated and hand corrected to ensure they define well-formed lambda-calculus expressions (for example, to ensure there are no cycles of anaphor or scope associations, and to ensure all variables in each elementary predication are bound by a lambda abstraction in the restrictor or nuclear scope of an outscoping quantifier). 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. This removes a source of error and simplifies their training.

Most quantifiers (including almost all quantifiers over eventualities) are existential and low-scoping (e.g. there is an eventuality for each combination of participants). Many articles also include multiple existentially quantified variables at the widest scope. Within each such group of quantifiers, their relative scope makes no truth-functional difference; these are left unannotated, and are assumed to be underspecified. Lambda calculus expressions are then generated by automatically and arbitrarily inducing scopes for

level of interaction	% of sentences
no interactions:	55.0
2 interacting quantifiers:	25.7
3 interacting quantifiers:	11.0
4 interacting quantifiers:	5.3
>4 interacting quantifiers:	3.0
total:	100.0

Table 2: Percent of sentences with 2, 3, 4, or more than 4 scopally interacting quantifiers annotated in the first 100 articles of the corpus.

these variables in a post-process. However, scoping evaluations described below are based on only the hand annotations, not the automatically induced scopes.

Table 2 shows the distribution of sentences with 2, 3, 4, or more than 4 scopally interacting quantifiers in the first 100 articles of the corpus (excluding article titles). Nearly half (45%) of all sentences show at least one interaction between a pair of quantifiers, and almost 20% show multiple interactions.

6. Inter-annotator Agreement

The corpus described in this paper is annotated for both coreference and scope, but inter-annotator agreement (IAA) of coreference is typically much higher than that of scope. Therefore, we here evaluate agreement on scoping.[9]

After the first 1,000 sentences were annotated, 33 articles of 3 sentences were sampled at random for a second, independent markup, IAA calculations, and error analysis.

6.1. Statistical Methods

The scopal relationship between any two quantifiers can be classified as direct, inverse, or non-interaction, and this classification has previously been the basis for IAA calculations. But with three or more quantifiers in an article, their relationships constrain one another, because scopings must be transitive and acyclic. This violates the independence assumptions of the de-facto standard κ statistics (Cohen, 1960; Davies and Fleiss, 1982), which are defined in terms of individual classifications rather than whole scopings, and so it invalidates their models of chance agreement. The granularity mismatch is particularly bad when there are many mutually constraining relationships, as in these multi-sentence articles.

For a more appropriate IAA statistic, we follow Artstein and Poesio (2008) and Skjærholt (2014) and adopt Krippendorff's α (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.

Freedom to select a distance metric addresses the problem of IAA over annotations with internal structure, such as scopings. The metric can compare annotations at the proper granularity (here, whole articles as opposed to single pairs of quantifiers), and in a way appropriate to their content. '[F]ine-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).

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.

6.1.1. Preprocessing

Preprocessing creates scope arcs that are implied by transitivity or by the scope of an anaphor's antecedent, a copula's complement, an appositive, etc. However, it does not generate arcs about which no two annotators could disagree, i.e. those indicating that the widest scope indirectly outscopes everything that it does not directly outscope, and those indicating that an existentially quantified eventuality is outscoped by its participants.

6.1.2. Correspondence

Two annotations are 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 articles.
3. If one article has more annotated scope-bearers than the other, top-scoped dummy entries fill out the correspondence.

Under this rule, two annotations of the same article will be matched mention-for-mention, and the top-scoping of dummy entries reflects the handling of top-scoped existentials in our annotation procedure. Annotations of different articles are matched up arbitrarily, to model chance agreement. The arbitrary matching simulates an annotator whose product has no relation to the contents of the article, but who does mark scoping and inheritance dependencies just as often as the real annotators do, and who shares their slight bias toward non-interacting existentials.

6.1.3. Symmetric Difference

Symmetric difference as a distance metric is motivated as follows: Truth-functionally necessary scopings represent a partial order over scope-bearers. The conventional distance metric between total orders is τ (Kendall, 1938), which is generalized to partial orders by Critchlow (2012). Critchlow's metric is computationally expensive (Brandenburg et al., 2012), but Malmi et al. (2015) prove an efficient

[9] Disagreements about reference may in fact entail disagreements about scope and thus be represented indirectly in our evaluation. However, we did not observe any in our error analysis.

approximation. With parameters appropriate for comparing scope annotations, the approximation is equivalent to the size of the symmetric difference between two graphs' arcs, which is itself a metric.

Size of symmetric difference is used for the 'constraint-level' κ of Manshadi et al. (2011), so our methodological explorations support their assessment of its value. However, the non-independence of individual scope arcs can inflate the agreement between two annotations. Krippendorff's α uses the simulated random annotator to reveal the extent of the inflation.

6.2. Findings

Chance-corrected $\alpha = 60.9\%$ using the distance metric just discussed. The metric takes account of agreement on two questions: For each pair of two scope-bearers in the same article, whether the scopal relationship between them is direct, inverse, or nonexistent/irrelevant to truth conditions; and for each individual scope-bearer, whether truth conditions require it to have any particular outscoper at all. For comparison, raw observed precision on these questions is 11,062 agreements out of 11,638 possible, or 95.1%.

Table 3 gives the inter-annotator confusion matrix for the scopal relationships of pairs of scope-bearers.

	direct	inverse	none
direct	24	6	35
inverse	6	128	140
none	47	192	10224

Table 3: Confusion matrix for scopal relationship.

Although these isolated pairs are not the proper granularity for chance correction, as discussed in Section 6.1., we offer measures calculated over them for the sake of comparison with existing resources.

Over only those pairs that both annotators labeled as interacting (i.e. *direct* or *inverse*), we obtained a Fleiss's κ of 0.755. Limiting the calculation to these pairs was motivated by a concern that apparent agreement would be inflated by the many pairs of scope-bearers with no truth-conditionally necessary relationship. In fact, though, when judgements of non-interaction were included, disagreements about the presence of scopal interaction lowered our κ to 0.409, versus 0.750 for the comparable measure in QuanText (Manshadi et al., 2011) or a Cohen's κ of 0.52 in the Higgins and Sadock (2003) data.

Table 4 gives the confusion matrix for a scope-bearer's needing to be outscoped to have the correct truth conditions.

	necessary	unnecessary
necessary	134	64
unnecessary	86	552

Table 4: Confusion matrix for necessity of outscoper.

6.3. Error Analysis

Annotator disagreements, sampled evenly from each cell of the confusion matrices, were traced to 47 root causes, summarized in Table 5.

trivial error	10
correspondence heuristic overused	10
guideline neglected	10
non-quantifier guideline lacking	6
other guideline lacking	6
different readings preferred	3
preprocessing software bug	2

Table 5: Reasons for annotator disagreement.

The *trivial error* class includes miswritten annotations and overlooked scope-bearers. Note that global disagreements (Table 3) are overwhelmingly about the presence of scopal relationships, not their direction. Most of these are from overlooking a scope-bearer. To limit this, annotators now receive their texts with placeholder annotations on each noun, and can search them for placeholders neither filled in nor deleted.

The *correspondence heuristic* is that correspondences between sets of objects are often scopal. Overreliance on the heuristic led to scope annotations among existentially quantified variables [10] that correspond to one another because of joint participation in an eventuality. We now have annotators regularly review their work with others, which helps to control non-trivial errors such as these, and we have rewritten our annotation guidelines for clarity.

Annotation guidelines were inadequate concerning non-quantifier scopal operators and in four other areas, leading to inconsistent personal judgements. Regular reviews help to identify and close these gaps.

7. Conclusion

This paper describes a corpus of 2,000 sentences from Simple English Wikipedia, which is intended to serve as training and evaluation data for broad-coverage quantifier scoping systems, and for the study of precise, complex sentence comprehension more generally. Nearly half of the sentences in this corpus are annotated with multiple scopally interacting quantifiers, suggesting the corpus is rich enough in quantifier scoping phenomena to serve as an evaluation dataset. Inter-annotator agreement suggests that the annotations in the corpus are reliable. This corpus is maintained at https://linguistics.osu.edu/schulerlab/dwnload.

8. Acknowledgements

This work was supported by National Science Foundation grant #1551313 and 1816891. All views expressed are those of the authors and not necessarily those of funders.

9. Bibliographical References

AnderBois, S., Brasoveanu, A., and Henderson, R. (2012). The pragmatics of quantifier scope: A corpus study. In

[10] Among which there can be no truth-functionally meaningful scoping.

Proceedings of Sinn und Bedeutung, volume 16, pages 15–28. MIT Working Papers in Linguistics.

Andrew, G. and MacCartney, B. (2004). Statistical resolution of scope ambiguity in natural language. *Unpublished manuscript*.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman, Harlow.

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.

Church, A. (1940). A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5(2):56–68.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher, editor, *The logic of decision and action*, pages 81–94. University of Pittsburgh Press, Pittsburgh.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

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.

Dinesh, N. (2010). *Regulatory conformance checking: Logic and logical form*. Ph.D. thesis, University of Pennsylvania.

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.

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.

Heim, I. (1982). The semantics of definite and indefinite NPs. *University of Massachusetts at Amherst dissertation*.

Higgins, D. and Sadock, J. M. (2003). A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96.

Kamp, H. (1981). A theory of truth and semantic representation. In Jeroen A. G. Groenendijk, et al., editors, *Formal Methods in the Study of Language: Mathematical Centre Tracts 135*, pages 277–322. Mathematical Center, Amsterdam.

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*.

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.

Leslie, S.-J. (2015). Generics oversimplified. *Nous*, 49(1):28–54.

Link, G. (1983). The Logical Analysis of Plurals and Mass Terms: A Lattice-theoretical Approach. In Paul Portner et al., editors, *Formal Semantics: The Essential Readings*, pages 127–147. Wiley-Blackwell.

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.

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.

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.

Parsons, T. (1990). *Events in the Semantics of English*. MIT Press.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Schuler, W. and Wheeler, A. (2014). Cognitive compositional semantics using continuation dependencies. In *Third Joint Conference on Lexical and Computational Semantics (*SEM'14)*.

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.

VanLehn, K. A. (1978). Determining the scope of English quantifiers. Technical Report AI-TR-98, MIT, Cambridge, Massachusetts.

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.