Argument & Computation 8 (2017) 113–129
DOI 10.3233/AAC-170020
IOS Press

Finding enthymemes in real-world texts: A feasibility study

Olesya Razuvayevskaya * and Simone Teufel
Computer Laboratory, University of Cambridge, Cambridge, United Kingdom
E-mails: or264@cam.ac.uk, sht25@cam.ac.uk

Abstract. Enthymeme reconstruction, i.e. the task of reformulating arguments with missing propositions, is an exciting task at the borderline of text understanding and argument interpretation. However, there is some doubt in the community about the feasibility of this task due to the wide range of possible reformulations that are open to humans. We therefore believe that research on how to define an objective ground truth for these tasks is necessary before any work on the automatic reconstruction can begin.

Here, we present a feasibility study for the task of finding and expanding enthymemes involving *a fortiori* arguments in real-world texts, and we show that given a sufficiently strict reformulation of the human annotation task, substantial agreement can be achieved. We split the task into three sub-tasks: 1. deciding whether a candidate text span really represents an enthymematic argument, 2. classifying the type of *a fortiori* argument concerned and 3. describing the missing premise in natural language. In a case study involving the two authors of this paper as annotators, we test a specific type of *a fortiori* arguments, the *let alone* construction, for its suitability for reaching high agreement in all three stages of the task. We also discuss pragmatic effects of *let alone* and how they relate to argumentation theory.

Keywords: Argumentation, enthymemes, annotation experiment, pragmatics, *a fortiori*

1. Introduction

According to the Aristotelian definition [6], *enthymemes* are standard-form syllogisms with one missing proposition. But in the modern usage of this term in argumentation theory, enthymemes are often defined as any type of arguments where one of the principal inferences is missing [2,7,12,45]. These missing inferences often express generally known facts. Consider the following argument:

(1) *All reptiles are cold-blooded animals. Therefore, lizards are cold-blooded.*

Example (1) is a syllogism where the inference from the first sentence to the second one, *lizards are reptiles*, is not stated explicitly as it is a widely known fact.

Enthymeme resolution is the task of restoring the original meaning of an argument. It often comprises a set of procedures, such as paraphrasing, reformulation, and insertion of generally accepted facts. The computational task of enthymeme reconstruction can be defined as the automatic insertion of missing premises or conclusions in naturally occurring enthymemes. It is an important step of the argument mining pipeline. Mining a collection of documents for arguments is a multistage process of identifying arguments in text, analysing smaller segments within these arguments and understanding relations between those segments. Peldszus and Stede [27] define five hierarchical steps of the task of argument

*Corresponding author. E-mail: or264@cam.ac.uk.

mining: *argument extraction*, *segmentation*, i.e. identification of minimal *argumentative discourse units* (ADUs), *segment classification*, i.e. labeling of ADUs based on their argumentative roles, *identification of relations* between these segments, and *argument completion*, i.e. automatic construction of statements from implicit propositions. The majority of currently existing computational techniques are aimed at the first four steps in this hierarchy, which are related to the analysis of explicitly specified argumentative information: detection of rhetoric structures [22,41] and argument zones [38–40], extraction of arguments [1,14,32] and detection of argumentative components [9,24]. Our goal is to contribute to the final stage, the problem of detecting and inserting implicitly stated argumentative propositions. It is a new prospect for text understanding and argumentation theory. The automatic reconstruction of enthymemes would be highly desirable for various applications, such as knowledge mining, i.e. the automated acquisition of common knowledge information that can potentially be stored in knowledge bases, argument validity verification, and automatic simplification of written texts for children or adult readers of lower reading age. It would also further our insights into the process of text understanding and argument interpretation, in particular into the question which inference steps are necessary when reading and understanding a text.

However, the literature on argumentation theory is generally pessimistic about the feasibility of an objective standard for enthymeme reconstruction [3,15,17]: the goal has been called a "challenging and subjective task even for human experts, often resulting in a wrong interpretation or exaggeration of the arguments" [34]. Hitchcock [17] mentions two particular problems associated with enthymemes, a) the difficulty of distinguishing enthymemes from deductively valid arguments and from arguments that are to be rejected ("demarcation problem") and b) the difficulty of evaluating whether the inferred expression was the one that was originally intended.

Due to the known inherent subjectivity of the enthymeme reconstruction task, a ground truth for this problem can in our opinion only be established by means of human annotation. If two or more annotators, working independently, can agree on the existence and identity of the missing premise, we will accept it as a proof that the enthymeme objectively "exists". This turns the problem of the subjectivity of enthymeme reconstruction into an empirical question, which can be answered simply by measuring agreement. We investigate the feasibility of several sub-tasks associated with this task:

- Agreement on whether the given argument is enthymematic, i.e., whether there exists a missing inference;
- Agreement on which subtype of inference is concerned;
- Agreement on the missing premise explicitly stated in natural language.

The annotation scheme constructed for this study is described in Section 4. In this work, we investigate a specific case of *a fortiori* arguments that can be interpreted without additional context and with a minimum amount of general knowledge. We discuss the properties of such arguments in Section 3.

The methodology and results of our experimental study are presented in Section 5. During the first step, we verify whether annotators agree on the existence of the missing proposition in a given text piece which is potentially representing an *a fortiori* argument. Annotators then select the category which best explains which type of proposition is missing from a fixed set of possible categories. The final step of reformulation of the missing premise is again performed under strict rules. We designed a linguistic template for each category; if a certain category is chosen, annotators produce their reformulation by inserting textual material from the sentence into the associated template. Within the constraints of this procedure, annotators are asked to create a reasonably naturally-sounding sentence.

This methodology attempts to make the task of enthymeme insertion more deterministic by constraining the variety of both reasoning steps and linguistic ways of formulating these steps so that:

(a) Minimal or no general knowledge and external information is needed for argument interpretation;
(b) There exists a unique logical reasoning scheme associated with the enthymeme concerned;
(c) The number of choices for formulating the missing reasoning step is restricted.

2. Properties of enthymemes

2.1. Enthymematic syllogisms

All types of arguments can be expressed in a truncated form, but truncated syllogisms are the most studied types of these [7]. Standard-form syllogisms are convenient for the analysis of enthymemes, because such arguments consist of only three terms and satisfy a set of strict requirements. These requirements uniquely define the form and the order of argumentative propositions and terms [36]. Therefore, given any two syllogistic propositions, the goal of restoring the missing premise or conclusion can be achieved by following a set of deterministic steps. For instance, consider the enthymeme below:

(2) *Tweety is a bird.* [minor premise]
Therefore she flies. [conclusion]

As can be inferred from the formal definition of a syllogism, this is a categorical syllogism with the *minor premise* and the *conclusion* expressed. There exists a general procedure [7] that allows us to conclude that the missing proposition is:

All birds fly. [major premise, universally quantified]

The procedure for restoring the missing proposition is the following:

(a) Define the three terms or categories associated with the syllogism. In Example (2), these terms are: *Tweety*, *bird* and *fly*.
(b) Classify the given terms into major, minor and middle ones. By the definition of a syllogism, the major term is the predicate of the conclusion, and the minor term is its subject. Therefore, *Tweety* is the minor term and *fly* is the major one. The remaining term, *bird*, is called the *middle* term. The middle term is shared by the major and the minor premises.
(c) Given the major and the minor terms, we can identify the types of the premises. The premise that contains the minor term as its subject is the minor premise. In our example, the first proposition is therefore the minor premise, and we can conclude that the missing proposition is the major one.
(d) Now, in order to construct the missing major premise, we must know that the major premise contains the major term as its predicate and the middle term as its subject. We therefore construct the following premise: *Birds fly*. Now, to make the syllogism valid, we have to make this premise universally true by means of applying the universal quantifier: *All birds fly*.

In this way, enthymematic syllogisms allow for a straight-forward and objective definition of the missing proposition: they already physically contain all three terms required for the reconstruction, so that no external information is necessary. This means that the process of reconstructing syllogistic enthymemes can potentially be automated. However, it is almost impossible to find well-formed standard syllogisms in everyday language, as we shall see later.

2.2. Enthymemes in everyday language

Unlike syllogisms, the majority of arguments used in natural language discourses are not formed from a set of structurally pre-defined propositions. Instead, people use a set of reasons to support their claims [29]. These reasons can be *non-defeasible*, i.e. universally true, or *defeasible*, i.e. they can be defeated by some new information [25,28]. The propositions left implicit in enthymemes are often defeasible, because the speakers prefer not to state potentially weak reasons explicitly [45]. For example:

(3) *Of course he does not know about differential equations. He is a lawyer.*

The missing proposition here, *all lawyers do not know about differential equations*, is clearly not universally true.

Information attacking defeasible reasons is called *defeater*. There exist two types of defeaters, *undercuts* and *rebuttals*. The undercutting defeater attacks the connection that exists between the reason and the conclusion, while the rebutting one attacks the conclusion itself. For example, recall the enthymeme mentioned in Example (2). It can be attacked by both the undercut, *Penguins are also birds, but they do not fly*, and the rebuttal, *But tweety is too heavy to fly*.

Due to the defeasible nature of argumentative claims used in everyday language, formal notations, such as syllogistic logic, cannot be applied for the restoration of naturally occurring enthymemes [10]. For instance, Campbell [4] argues that the majority of Darwin's claims cannot be expressed through logical arguments. The mechanical approach of reconstructing enthymemes in such a way that they become structurally correct and include the needed assumptions may lead to incorrect interpretations of an argument, i.e. to the insertion of premises not intended by the speaker. In Example (2), *All birds fly* exemplifies this problem. The statement intended by the speaker might be defeasible: *Generally, birds fly*. Therefore, our attempt to convert the statement into a well-formed syllogism by means of the universal quantifier results in an incorrect statement.

Implicit argumentative propositions are not always defeasible. Some of them represent universally true facts that are too trivial to be stated explicitly [18]. This observation is captured by Grice's Maxim of Quantity – *Do not make your contribution more informative than is required* [16]. This pragmatic principle discourages a speaker from making explicit statement of facts that are widely accepted by the audience; in fact, the explicit statement of such facts would sound both trivial and also unnatural to a listener. This is another reason why enthymemes are so frequent in natural language discourses.

Jackson and Jacobs [19] notice that enthymemes also play an important role in maximizing a listener's agreement, because additional information may contain a subjective point of view and increase the possibility of disagreement. For example:

(4) *There is no law against composing music when one has no ideas whatsoever. The music of Wagner, therefore, is perfectly legal.* Mark Twain

The missing proposition in this humorous argument, *Wagner composed music while having no ideas whatsoever*, is subjective, and many people will disagree with it. The author therefore refrained from stating it explicitly.

The detection of hidden and implicitly stated arguments is much harder than the treatment necessary for enthymematic syllogisms; it would require a thorough analysis of argumentative structures and strong general or domain-specific knowledge. The question of how much domain knowledge is needed becomes particularly critical when enthymemes appear in scientific domains; their resolution is therefore usually reviewed as an activity for experts in the area [4]. Crick states that enthymemes are frequent

in scientific writing, as a result of the fact that the audience is assumed to consist of experts in the field [8]. Enthymemes make scientific persuasion more engaging and initiate cognitive cooperation between the writer and the audience, hence creating a stronger persuasion effect than if complete arguments had been used, which may be initially surprising. According to Campbell, the effect is achieved by the shared reasoning process between writer and audience, where missing premises are jointly filled in [4]. Because of their persuasive power, Aristotle called enthymemes "the strongest of rhetorical proofs".

As can be seen from the reasons that explain the occurrence of enthymemes in everyday language, the task of restoring the assumptions intended by the speaker is an unfeasibly challenging task, particularly if one does not have access to the formal structure of the argument. It relies on general knowledge and on subjective interpretation of an enthymeme, given the unlimited number of choices and humans' viewpoints [10]. Moreover, it is not clear how many assumptions should be inserted in order to make the argument complete. Therefore, our solution relies on a fixed argumentation strategy, thereby carefully restraining the theoretically possibly unlimited choice. As a result, we do not require a fully mechanical procedure for enthymeme reconstruction.

3. *A fortiori* arguments

In this study, we concentrate of *a fortiori* arguments, which were first mentioned in Aristotle's Rhetoric [6]. These arguments can be classified into two main groups, positive and negative arguments.

The main principle behind negative *a fortiori* arguments is that we compare two cases, one where a certain quality is more likely to exist and some other case where it is less likely to exist. If we now assert that even in the more likely case, the quality does not exist, the logical conclusion is that the quality certainly cannot exist in the less likely case. The *a fortiori* argument therefore serves as a refutation of the less likely case. This type of *a fortiori* argument is called *a minore ad maius*, i.e., from smaller to bigger.

An equivalent case called *a maiore ad minus* exists for positive *a fortiori* reasoning. This is based on the inference about the existence of the quality for the more likely case, based on the fact that even the less likely case has this quality.

3.1. Kienpointner's scheme and extension

Kienpointner [20] designed the following argumentation scheme for *a minore ad maius* arguments:

> If even X does not have property P and it is a less likely case that Y has property P than that X has property P, then Y does not have P
> (Even) X does not have P
> (Therefore) Y does not have P

We formulate the equivalent scheme for *a maiore ad minus*:

> If even X has property P and it is a more likely case that Y has property P than that X has property P, then Y has P
> (Even) X has P
> (Therefore) Y has P

Our observation is that while this scheme is correct and explanatory, it does not express the additional implicit scaling relation that holds between X and Y. Consider the following example:

(5) *If even the house could not accommodate them, the room definitely will not.*

Kienpointer's scheme only allows expressing:

> *If even the house does not have a property of accommodating them, and it is a less likely case that the room has a property of accommodating them than the house has the property of accommodating them, then the room does not have a property of accommodating them.*

As can be seen, the suggested insertion does not explain the key scaling relation *size* between X and Y:

> *The room is smaller than the house.*

To address this problem, we extend Kienpointer's argumentation scheme with a fourth proposition that describes a concrete relation holding between X and Y:

> *X and Y stand in Relation R.*

The relation R between X and Y is almost always implicit, which makes *a fortiori* arguments ideal cases for our investigation of the feasibility of reaching agreement in an enthymeme interpretation task.

3.2. Nedum and let alone sentences

A fortiori arguments can be expressed in natural language by many linguistic constructions, involving for instance *of course, not...even*, and *if...then*.

(6) My child *can* already *write*. **Of course** he *knows the alphabet*.

(7) I could **not** run **even** *3 miles*. I definitely will not be able to run this *marathon*.

(8) **If** *John* could solve this task, **then** *you* for sure will.

In this study, we investigate one of the most typical *a fortiori* constructions: *let alone* sentences. *Let alone* is a convenient *a fortiori* construction, because all the necessary components are usually present in a single sentence, and the amount of argument-irrelevant information in that sentence is minimal. Consider Example (5) paraphrased for *let alone* construction:

(9) *Even the house could not accommodate them, let alone the room.*

Let alone is a *nedum* – a connective particle that signalizes that the second clause of the construction is more relevant to the context and can be inferred from the first clause [33]. For example:

(10) *This task is difficult for an **adult**, let alone a **child**.*

(11) *The baby cannot **sit** yet, let alone **walk**.*

According to Fillmore et al. [13], *let alone* can be syntactically treated as a coordinating conjunction, where the phrases linked by *let alone* display parallelism with respect to grammatical functions, e.g. direct object (*adult* and *child*) or verbal complement to auxiliary (*sit* and *walk*). In the terminology of Toosarvandani [42], the stressed elements in the first and second part of the sentence are called *correlate* and *remnant* respectively.

Let alone can be used to signal both positive and negative *a fortiori* arguments, representing *a minore ad maius* and *a maiore ad minus* arguments respectively. Sawada [31] distinguishes the following three main cases of *let alone* arguments: explicit negative (Example (12)), implicit negative (Example (13)),

and explicit positive (Example (14)):

(12) *He cannot solve simple, let alone starred tasks.*

(13) *The task is too difficult to be understood, let alone solved.*

(14) *He could find the exact answer, let alone an approximate solution.*

Toosarvandani [42] observes that some form of scaling information is always present between the remnant and the correlate, making the correlate in some respect more likely for negative sentences and less likely for positive ones. The hearer can pragmatically infer, based on common knowledge [16,35], that these stressed elements are ordered based on some scale. This hidden scalar reasoning is highly relevant to our task of finding arguments with hidden premises, and is captured in the extension to Kienpointner's scheme presented in Section 3.1 above. We argue that *let alone* sentences are by definition enthymemes, and that the missing scalar relation, called relation R above, is an important part of the missing proposition necessary for the conclusion to be accepted. The conclusion is therefore always a statement about the remnant, which is entailed from both the parallel statement about the correlate and the inserted scalar relation.

4. Annotation scheme

We categorised the scaling relations based on a development corpus of about 250 cases of *let alone* sentences, resulting in the scheme in Table 1.[1]

Table 1
Annotation scheme for *let alone* scales

Category	
1	Part of
2	Smaller than
3	Precondition for
4	Other lexical entailment
5	Earlier date
6	Additional constraint
7	Additional referents
8	Cumulative/independent
9	More extreme case than
10	Easier than
11	Less likely than

The first 8 categories in our scheme were motivated by the most frequent *let alone* cases observed in the development corpus. These categories are neither mutually exclusive nor are they at the same level, but they are in our opinion the ones that characterize the various phenomena best. For example, any pair of remnant and correlate standing in relation *part of* will by default also satisfy the *smaller than* relation. The general annotation rule therefore was to select the highest-priority category that describes the situation. We define the priority by specificity and order the categories from more specific to less

[1] The agreement reported subsequently in this article is measured on a separate test corpus.

specific. As *part of* is more specific than *smaller than*, the annotators would have to give it preference. For the types of scalar relations that rely on very subjective inferences or the ones that were not observed in the development set but that might occur during testing time, we have identified the three fallback categories 9–11, which are more generic than the main ones.[2]

We will now describe the categories.

4.1. Category definition

1. Part of: The relationship between the referents in remnant and correlate is that of superset. For example:

(15) *This does not apply to Germany, let alone Europe.*

The scaling relation R here is: *Germany* is a part of *Europe*.

This relation has a number of requirements:

(1) The statement holds for the holonym (whole) only if it holds for all of its meronyms (components).
(2) The meronym does not have any property that would make it an exceptional case in relation to the rest of the cases (as this is covered in category *Additional constraint* below). For example, if we know that Germany is much less likely to accept a particular law than the rest of the Europe, we could make a different argument from superset to the subset:

(16) *No country in Europe will accept this law, let alone Germany.*

This case is not covered here, but in the category called *Additional constraint*.

2. Smaller than: The scale concerned involves cardinalities or standard measurements which connect the remnant and the correlate. For example:

(17) *You wouldn't make to New York, let alone the West Coast.*

(18) *There is no space in the fridge for a bottle of water, let alone a saucepan.*

The scaling relations here would be:

The distance to New York is smaller than *the distance to the West Coast*

A bottle of water is smaller than *a saucepan*

Such arguments compare measurements of the correlate and remnant's properties (e.g., weight, age, volume). In an Example (18), the scalar comparison is between the volumes of *a bottle of water* and *a saucepan*:

The volume of a bottle of water is smaller than *the volume of a saucepan*

However, in order to make the annotation task more straightforward, we omit the obligatory mention of the property and accept direct scalar comparisons between the remnant and the correlate.

[2]Our choice of three categories rather than one was motivated by our wish to create naturally-sounding template sentences. However, in future work, we will choose only one fallback category, at the cost of somewhat less elegant template sentences.

3. Precondition: The action in the remnant necessarily requires the action in the correlate having taken place earlier. For example:

(19) *Your talent isn't enough to participate, let alone win.*

The scaling relation here would be: *participating* is a precondition for *winning*.

Fellbaum [11] presents four types of entailment relations between verbs; our *precondition* corresponds to her *backward presupposition* (*winning* presupposes *participating*).

4. Other lexical entailment: This relation covers the larger class of actions where the proposition expressed in the remnant lexically entails the proposition expressed in the correlate. There can be temporal overlap between the two propositions, unlike in category 3, where the proposition in the correlate has to be completed before the proposition in remnant can begin.[3] For example:

(20) *He doesn't even sleep, let alone snore.*

The scaling relation here would be: *snoring* entails *sleeping*. *Snoring* can only happen during *sleeping*. Additionally, *snoring* and *sleeping* can occur simultaneously for some of the time, but it is not necessary that *snoring* entirely overlaps with *sleeping*; there can be times of *sleeping* without *snoring*.

5. Earlier date: This is a special case of *Smaller than* category, where the property is time: in the case of irreversible events, if a state becomes true at time t_0 for the first time, by definition it does not hold at any earlier time $t < t_0$. This relation relies on the uni-directionality of this particular scale, time. For example:

(21) *They didn't have electricity in 1923, let alone 1909.*

The scaling relation here would be: *1909* is an earlier date than *1923*.

6. Additional constraint: In his category, a number of semantic predicates (constraints) applied to the correlate are compared to the same set of predicates plus additional predicates in the remnant. For example:

(22) *You don't know what a middle-aged person feels like, let alone a middle-aged prince.*

Here, a *middle-aged prince* is a double-constraint on a person, which filters out some middle-aged non-princes.

7. Additional referents: This category covers cases where the set of referents in the remnant logically includes the referent from the correlate. For example:

(23) *The company does not even insure their employees, let alone their families.*

[3] This is analogous to the union of Fellbaum's categories *co-extensiveness* and *proper temporal inclusion*. *Co-extensiveness* of events holds in one direction for the entire time period (*moving* always happens whenever *walking* is taking place). In contrast, *proper temporal inclusion* may hold between two events that happen to be performed simultaneously, but it is not necessary that both overlap for the entire duration of one of the actions. For instance, *snoring* is possible during some *sleeping* interval. The distinction between these classes is known to be difficult even for humans [11,43], and of no consequence to our task, so we collapsed them.

It is hard to imagine a company that would insure the family of an employee without also insuring the employee himself. The intended meaning of the remnant must therefore be logically interpreted as "the employee and his family".

8. Cumulative/independent: Here, the remnant of *let alone* is not directly comparable to the correlate, but becomes interpretable if we read it as an additive constraint. For example:

(24) *This Easter-egg packaging does not even protect its contents, let alone have anything to do with Easter.*

Here, both the feature of protecting the contents and of being decorated with Easter motives are important for Easter egg packaging. The intended meaning that speaker put into this argument is:

This Easter-egg packaging does not even protect its contents. In addition, it does not have anything to do with Easter.

Hence, the combination of the two qualities of the Easter-egg packaging is compared to the quality of protecting its contents. One possible test for whether the relation is *independent/cumulative* is to try to substitute *let alone* with *in addition to*. If the meaning does not change, then the remnant and the correlate are likely to be in the *independent/cumulative* relation.

The three categories **Easier than**, **Less likely than**, and **More extreme case than** are fallback options, which are only to be used if none of the more specific categories applies. For example:

(25) *They refused to refer to Kursk, let alone Moscow.*

(26) *I could not solve the first, let alone the last tasks.*

(27) *I have not even seen Mary, let alone Rose.*

4.2. Linguistic templates

There is a connection between the scale category introduced in the previous section and the linguistic form of premise that can be constructed from that class, as can be seen from Table 2.

Table 2
Linguistic templates

Category		Linguistic template
1	Part of	X is a part of Y
2	Smaller than	X is smaller than Y
3	Precondition for	X is a precondition for Y
4	Other lexical entailment	Y lexically entails X
5	Earlier date	Y is earlier than X
6	Additional constraint	Y poses additional constraint on X
7	Additional referents	Y in addition contains X
8	Cumulative/independent	X and Y are cumulatively more important than X
9	More extreme case than	Y is a more extreme case than X
10	Easier than	X is easier than Y
11	Less likely than	Y is less likely than X

Each scaling relation is represented by one template out of the many possible verbalizations of that relation. Annotators were asked to extract the appropriate parts from the given sentence and insert them into the template of their chosen category. For instance, for Examples (15)–(27) they were expected to write the sentences in Table 3. (Material directly excised from the sentence is shown in boldface.)

As can be seen from these examples, some of the statements thus constructed are always true, such as: *Snoring lexically entails sleeping*, *Germany is a part of Europe* or *participating is a precondition for winning*. However, some of the statements are true only in the given context. For example, *The distance to New York* is smaller than *the distance to the West Coast* holds only for certain speaker locations. Nevertheless, such statements can be good candidates if our aim is to represent knowledge about generally accepted facts.

Table 3
Filled linguistic templates for categories from Table 1

Category		Verbalized statement
1	Part of	**Germany** is a part of **Europe**.
2	Smaller than	The **distance to New York** is smaller than the **distance to the West Coast**.
3	Precondition for	**Participating** is a precondition for **winning**.
4	Other lexical entailment	**Snoring** lexically entails **sleeping**.
5	Earlier date	**1909** is earlier than **1923**.
6	Additional constraint	**Being a middle aged prince** poses additional constraints on **being a middle-aged person**.
7	Additional referents	The **families of employees** in addition contain **employees**.
8	Cumulative/independent	**Protecting the contents** and **having relation to Easter** are cumulatively more important than **protecting the contents**.
9	More extreme case than	**Moscow** is a more extreme case than **Kursk**.
10	Easier than	The **first task** is easier than the **last task**.
11	Less likely than	**Seeing Rose there** is less likely than **seeing Mary there**.

5. Annotation experiment

100 random *let alone* sentences were extracted from the British National Corpus (BNC). The annotation experiment consisted of three sub-tasks:

(a) Classification of the given sentences into enthymemes or non-enthymemes;
(b) Classification of the relation type holding between the remnant and the correlate (as described in Section 4 above);
(c) Filling the corresponding linguistic template with linguistic material (as described in Section 4 above).

5.1. Enthymeme classification task and results

Two annotators (the authors of this paper) made a binary decision when faced with a *let alone* sentence; they had to decide whether the sentence contained enough context to interpret the sentence and to restore the missing statement. They additionally crossed out any information that was not relevant to the argument expressed in the *let alone* sentence, or actively distracting.

The number of positive choices was high (183 out of 200). The inter-rater agreement was calculated by means of Cohen's *kappa* statistics [5], to be $K = 0.729$ ($N = 100$, $k = 2$, $n = 2$). We use Cohen's standard formula:

$$k = \frac{p_o - p_e}{1 - p_e}$$

p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. The numerator represents the difference between the observed probability of success and the probability of success under the assumption of an extremely bad case. Independence implies that a pair of raters agree about as often as two people who effectively flip coins to make their ratings. Kappa's possible values are constrained to the interval $[-1; 1]$. $K < 0$ means that the observed agreement is less than the one expected by chance, $K = 0$ means that agreement is not different from chance, and $K = 1$ means perfect agreement. However, just obtaining kappa significantly greater than zero is not sufficient to evaluate the quality of the agreement. Various schemes to assess kappa's significance have been proposed [30,44]. According to the strictest scale proposed by Krippendorff, our annotation was "marginally reliable" and well within the range of what is accepted in Computational Linguistics as good agreement for pragmatics and discourse-level annotation [21].

There were only 2 out of 183 cases where either annotator specified enthymeme-irrelevant information, namely:

(28) ~~*According to Rex,*~~ *we never had a chance of negotiating the bend, let alone stopping, given how fast we were travelling.*

(29) *'But the king is not yet married, let alone having a son,'* ~~*objected Anne – but then her face brightened.*~~

5.2. Enthymeme classification task

The next stage of the experiment is performed only on those *let alone* sentences that were positively classified by both annotators as enthymematic (90 cases).

The annotators classified the *let alone* sentences according to the annotation scheme from Table 1 in order to describe the scale that applies for the *a fortiori* argument.

The agreement between annotators was $K = 0.75$ ($k = 2$, $n = 2$, $N = 90$). This indicates that annotators were overall well able to distinguish scaling relation categories.

The class distribution per category is skewed, as can be seen in Table 4. The most frequently chosen category was the "precondition for" relation – 61 cases (33.9%). The "additional constraint" and "additional referent" classes were also frequent (12.7% and 10% respectively). The fact that the fallback categories were chosen relatively frequently suggests that there are cases where the context is not enough to disambiguate the scale.

The small sample size prevents us from analyzing problematic categories systematically, but we noticed that *earlier date* and *precondition for* seem to be relatively harder for the annotators to distinguish reliably. Despite the very distinctive nature of the "cumulative/independent" category, the annotators frequently disagreed on this relation type. We believe that most of the disagreements for the fallback categories came from the fact that one of the annotators preferred these generic relations to more specific ones. For example:

(30) *They couldn't organise a strike, let alone a revolution.*

While one of the annotators classified this relation as *a strike is **smaller than** a revolution*, the second participant preferred the *easier than* category (*a strike is **easier than** a revolution*).

While such unclear cases are in the nature of annotating subjective interpretative phenomena, the relatively high agreement we observe means that overall they are rare.

Table 4
Class distribution (absolute and relative) per relation category

Relation type	Number of choices	
Part of	8	(4.4%)
Smaller than	9	(5.0%)
Precondition for	61	(33.9%)
Other lexical entailment with	11	(6.0%)
Earlier date	5	(2.7%)
Additional constraint on	23	(12.7%)
Additional referent	18	(10.0%)
Cumulative/independent	9	(5.0%)
Less likely than	9	(5.0%)
More extreme case than	12	(6.7%)
Easier than	15	(8.3%)

5.3. Enthymeme reformulation task

The annotators next composed a new sentence expressing the missing premise, using the textual material extracted from the given sentence together with the template for the chosen category, as described in Section 4. The 71 cases where the annotators previously agreed on the chosen scale category were used for this measurement.

We report the percentage of identical or near-identical statements. Judgments of what is identical were made by the first author, but they were very strict, allowing variation only with respect to near-synonyms or the addition or omission of non-relevant information.

69 premises (97%) were judged to be identical or near-identical.

Only two disagreements occurred, the first of which was:

(31) *Few adherents to the new classical macroeconomics trouble even to question it, let alone provide an analytical basis to justify it.*

While both annotators identified that "questioning" is a precondition for "providing an analytical basis to justify", they specified partially different propositions:

(a) *Questioning something is a precondition for providing an analytical basis to justify something.*

(b) *Questioning classical macroeconomics is a precondition for providing an analytical basis to justify classical macroeconomics.*

The second enthymeme on which the annotators disagreed was:

(32) *It's not exactly as if Wimbledon are an English force – let alone European.*

Here, the annotators disagreed on the ellipsis resolution:

(a) *For Wimbledon, to be an English force is easier than to be a European force.*
(b) *To be an English force is easier than to be a European force.*

Overall, two disagreement cases out of 71 is a surprisingly low number, given that we demanded linguistic descriptions of quite complex pragmatic effects, and given that the annotators, while knowing the scheme and the ideas, were still working entirely independently on a task that was hitherto deemed too unrestricted to be done at all. The pilot experiment in our opinion shows that applying the right kind of restriction can result in an objective definition of truth, even in seemingly open-ended situations.

6. Discussion and conclusions

In this work, we presented the first study on whether it is possible for humans to reliably find and reconstruct enthymemes in unrestricted texts. The study is limited in size and uses the two authors as annotators, but we nevertheless hope to have demonstrated that the *a fortiori* reasoning exemplified by the *let alone* construction is well-defined and can be intuitively described using human-directed rules. This is encouraging, particularly as the community was previously generally pessimistic about the feasibility of the enthymeme reinsertion task.

Another insight that this study suggests is the fact that pragmatic information, which is not traditionally used in argument mining, plays an important role in the resolution of this type of argument. From a practical viewpoint, we are also satisfied with the result of the study, because a reliable gold standard is required as evaluation material for the next stages of automatic enthymeme recognition.

For the case of *let alone*, we have observed high inter-annotator agreement on the decision of whether potentially enthymemic text units are indeed enthymemes. We further investigated whether the annotators also agree on the exact type of pragmatic relation that holds between two propositions. The results show that although there exist some types of relations that are not easily distinguishable from each other, and some that do lack context for interpretation, the annotators are overall able to achieve significant agreement on this task. In a third task of premise reformulation, the agreement was near-perfect. This result suggests that *let alone* constructions are ideal candidates for the task of detecting naturally occurring enthymemes in texts and reconstructing the missing propositions.

Let alone sentences are particularly convenient for the analysis of enthymemes because the information we need for interpreting the *a fortiori* argument is very localized, i.e., both the conclusion and the premise are usually located in one sentence. This reduces the amount of argument-irrelevant information. However, in the general case we would expect that other linguistic structures, such as arguments expressed with *because*, *of course*, *therefore*, occupy much larger text spans, resulting in a larger amount of argument-irrelevant information. For these more complicated constructions, human agreement is unlikely to be as high.

7. Future work

In future work, we plan to perform this task on a larger scale, with independent annotators and on a larger dataset. Despite the substantial inter-annotator agreement observed in this study, we plan to critically re-consider the relation types in the light of the annotation results. Some categories seem less clearly defined than others; for instance, the *part of* relation was often confused with the *additional constraints* relation. These two categories can be merged into one in the future, considering that they both describe a relation from general to particular. Also, the three fallback categories are very similar

to each other. Our original reason for differentiating them was to ensure that the respective templates sound maximally natural. But weighted against the redundancy contained in three functionally identical categories, we will in the future opt for a single fallback category.

A different improvement concerns the explanatory power of the scheme. As has been mentioned, scalar relations usually deal with the degree to which the remnant and the correlate have a particular property. Although we omitted the notion of a property in this annotation scheme, we consider strengthening its role and introducing it into the template, because being able to capture the property would provide more explanatory power concerning the factors affecting the interpretation of arguments. Another motivation for asking for an explicit statement of the property is its connection to the positive or negative version of the *a fortiori* argument. Our future annotation instructions should reflect the difference between the versions and the role of negation, possibly incorporating separate templates for positive and negative inferences.

As a next step in our research, we plan to expand the range of discourse markers used for *a fortiori* constructions beyond *let alone*, but we suspect that not all possible linguistic constructions are plausible candidates for enthymeme resolution. Other cases may not be as context-minimal as *let alone* sentences, i.e. the reason supporting the claim may often be located further away from the sentence containing the claim. This will add to the amount of argument-irrelevant information that has to be considered, which is likely to reduce the agreement in enthymeme interpretation. Moreover, other discourse markers may be less *a fortiori*-specific, i.e., they may be used to express different logical patterns. Consider the following two cases:

A must be true, because A has never been observed to be false.
A must be true, because X has told so.

While the same discourse marker is used in both cases, the corresponding enthymematic arguments are based on entirely different reasoning types. The first argument is based on the lack of evidence about any controversial case, whereas the second argument is based on the reliability of the claim's source. These arguments also correspond to different argument schemes, *argument from expert opinion* and *argument from ignorance*, and it seems plausible to us that different reconstruction techniques might be successful in the two cases. We believe that linguistic constructions that uniquely identify their associated argument schemes will result in a more objective enthymeme resolution task. It is therefore important to investigate a wide range of other linguistic structures in order to find other *a fortiori* behaviour similar to *nedum*.

Other things are likely to get more difficult too when moving beyond *let alone*. For instance, in the current experiment, rephrasings of the relations were often identical or near-identical, and thus so easy to judge objectively that we decided that one of us could make this judgement without introducing bias. With more challenging discourse markers, this may not always be the case. We therefore intend to outsource the decision of whether two paraphrases are identical or near-identical to unrelated humans, e.g., on Mechanical Turk [26], thus eliminating any possible bias.

Human annotation is only the first step towards automatic reconstruction of enthymemes. In particular, the automatic resolution of gapping is a challenging task. Syntactic information is often not enough for intelligent reconstruction, and the semantic information required is not available automatically [23,37]. We will investigate machine learning from syntactic structure and semantic information in the future. In this context, we predict that pragmatic features, if correctly modeled, should in the future beneficially influence the machine learning of argument structure.

References

[1] M.A. Angrosh, S. Cranefield and N. Stanger, Ontology-based modelling of related work sections in research articles: Using crfs for developing semantic data based information retrieval systems, in: *Proceedings of the 6th International Conference on Semantic Systems*, ACM, 2010, p. 14.

[2] E. Black and A. Hunter, Using enthymemes in an inquiry dialogue system, *International Foundation for Autonomous Agents and Multiagent Systems* **1** (2008), 437–444. ISBN 9780981738109.

[3] M.B. Burke, Unstated premises, *Informal Logic* **7**(2) (1985). doi:10.22329/il.v7i2.2709.

[4] J.A. Campbell, Scientific discovery and rhetorical invention: The path to Darwin's origin, *The Rhetorical Turn: Invention and Persuasion in the Conduct of Inquiry* (1990), 58–90.

[5] J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1) (1960), 37–46, http://epm.sagepub.com/content/20/1/37.short. doi:10.1177/001316446002000104.

[6] E.M. Cope and J.E. Sandys (eds), *Aristotle: Rhetoric*, Vol. 1, Cambridge University Press, 2010, Cambridge Books Online. URL, http://dx.doi.org/10.1017/CBO9780511707421. ISBN 9780511707421.

[7] I.M. Copi and C. Cohen, *Introduction to Logic*, Maxwell Macmillan International Editions, 1990, URL, https://books.google.co.uk/books?id=UnIbAQAAMAAJ. ISBN 9780023250354.

[8] N. Crick, Conquering our imagination: Thought experiments and enthymemes in scientific argument, *Philosophy and Rhetoric* **37**(1) (2004), 21–41. doi:10.1353/par.2004.0009.

[9] J. Eckle-Kohler, R. Kluge and I. Gurevych, On the role of discourse markers for discriminating claims and premises in argumentative discourse, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, September 2015, Association for Computational Linguistics, 2015, pp. 2249–2255.

[10] R.H. Ennis, Identifying implicit assumptions, *Synthese* **51**(1) (1982), 61–86. doi:10.1007/BF00413849.

[11] C. Fellbaum, *WordNet*, Springer, Dordrecht, Netherlands, 2010, pp. 231–243, ISBN 978-90-481-8847-5. doi:10.1007/978-90-481-8847-5_10.

[12] V.W. Feng and G. Hirst, Classifying arguments by scheme, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1, HLT '11*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 987–996, URL, http://dl.acm.org/citation.cfm?id=2002472.2002597. ISBN 978-1-932432-87-9.

[13] C.J. Fillmore, P. Kay and M.C. O'Connor, Regularity and idiomaticity in grammatical constructions: The case of let alone, *Language* **64**(3) (1988), 501–538, ISSN 00978507. doi:10.2307/414531.

[14] T. Goudas, C. Louizos, G. Petasis and V. Karkaletsis, Argument extraction from news, blogs, and social media, in: *Hellenic Conference on Artificial Intelligence*, Springer, 2014, pp. 287–299.

[15] J. Gough and C. Tindale, *Hidden', or 'Missing' Premises*, 1985. URL, http://scholar.uwindsor.ca/philosophypub/20.

[16] H.P. Grice, Logic and conversation, in: *Syntax and Semantics: Vol. 3: Speech Acts*, P. Cole and J.L. Morgan, eds, Academic Press, San Diego, CA, 1975, pp. 41–58.

[17] D. Hitchcock, *Enthymematic Arguments*, 1985.

[18] P.J. Hurley, *A Concise Introduction to Logic*. Cengage Learning, 2014, URL, https://books.google.co.uk/books?id=qGBQAwAAQBAJ. ISBN 9781285965567.

[19] S. Jackson and S. Jacobs, Structure of conversational argument: Pragmatic bases for the enthymeme, *Quarterly Journal of Speech* **66**(3) (1980), 251–265, ISSN 0033-5630. doi:10.1080/00335638009383524.

[20] M. Kienpointner, *Alltagslogik: Struktur und Funktion Von Argumentationsmustern*, Problemata (Stuttgart). Frommann-Holzboog, 1992, URL, https://books.google.co.uk/books?id=CsJlQgAACAAJ. ISBN 9783772814624.

[21] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage, 2004.

[22] W.C. Mann and S.A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, *Text-Interdisciplinary Journal for the Study of Discourse* **8**(3) (1988), 243–281. doi:10.1515/text.1.1988.8.3.243.

[23] M. McShane and P. Babkin, Automatic ellipsis resolution: Recovering covert information from text, 2015, http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9441.

[24] M.-F. Moens, E. Boiy, R.M. Palau and C. Reed, Automatic detection of arguments in legal texts, in: *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ACM, 2007, pp. 225–230.

[25] D. Nute, Defeasible reasoning: A philosophical analysis in prolog, in: *Aspects of Artificial Intelligence*, Springer, 1988, pp. 251–288. doi:10.1007/978-94-009-2699-8_9.

[26] G. Paolacci, J. Chandler and P.G. Ipeirotis, Running experiments on Amazon mechanical turk, *Judgment and Decision Making* **5**(5) (2010), 411–419.

[27] A. Peldszus and M. Stede, From argument diagrams to argumentation mining in texts: A survey, *Int. J. Cogn. Inform. Nat. Intell.* **7**(1) (2013), 1–31. doi:10.4018/jcini.2013010101.

[28] J.L. Pollock, Defeasible reasoning, *Cognitive Science* **11**(4) (1987), 481–518. doi:10.1207/s15516709cog1104_4.

[29] J.L. Pollock, A theory of defeasible reasoning, *International Journal of Intelligent Systems* **6**(1) (1991), 33–54. doi:10.1002/int.4550060103.

[30] T. Rietveld and R. Van Hout, *Statistical Techniques for the Study of Language Behaviour*, Mouton de Gruyter, Berlijn, 1993.
[31] O. Sawada, Rethinking the let alone construction: what are its construction specific characteristics, 2003.
[32] J. Schneider and A. Wyner, Identifying consumers' arguments in text, in: *SWAIE*, 2012, pp. 31–42.
[33] J. Schrickx, Nedum: 'much less' or 'much more'?, *Journal of Latin Linguistics* **15**(1) (2016), 117–144. doi:10.1515/joll-2016-0001.
[34] M. Scriven, Reasoning, 1977.
[35] R. Stalnaker, Common ground, *Linguistics and Philosophy* **25**(5–6) (2002), 701–721. doi:10.1023/A:1020867916902.
[36] J.Z. Sukkarieh, Mind your language! controlled language for inference purposes, 2003.
[37] N. Suszczańska, J. Romaniuk and P. Szmal, Automatic analysis of elliptic sentences in the thetos system1, 2005.
[38] S. Teufel et al., Argumentative zoning: Information extraction from scientific text, PhD thesis, Citeseer, 2000.
[39] S. Teufel and M. Moens, What's yours and what's mine: Determining intellectual attribution in scientific text, in: *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, Vol. 13, Association for Computational Linguistics, 2000, pp. 9–17.
[40] S. Teufel and M. Moens, Summarizing scientific articles: Experiments with relevance and rhetorical status, *Computational Linguistics* **28**(4) (2002), 409–445. doi:10.1162/089120102762671936.
[41] S.A. Thompson and W.C. Mann, Rhetorical structure theory: A frarnework for the analysis of texts, 1987.
[42] M. Toosarvandani, Letting negative polarity alone for let alone, in: *Proceedings from Semantics and Linguistic Theory XVIII*, T. Friedman and S. Ito, eds, CLC Publications, Ithaca, New York, 2008, pp. 729–746.
[43] G. Tremper, A. Frank, H. Zinsmeister and B. Webber, A discriminative analysis of fine-grained semantic relations including presupposition: Annotation and classification.
[44] A.J. Viera, J.M. Garrett et al., Understanding interobserver agreement: The kappa statistic, *Fam Med* **37**(5) (2005), 360–363.
[45] D. Walton and C.A. Reed, Argumentation schemes and enthymemes, *Synthese* **145**(3) (2005), 339–370.