Computational Models of Argument
C. Reed et al. (Eds.)

doi:10.3233/FAIA240328

Defining Argumentative Discourse Units as Clauses: Psycholinguistic Evidence

Clara SEYFRIED [ab,1], Chris REED [a] and Yuki KAMIDE [b]

[a] *Centre for Argument Technology, University of Dundee, UK*
[b] *Division of Psychology, University of Dundee, UK*

ORCiD ID: Clara Seyfried https://orcid.org/0000-0001-6933-2343

Abstract. Identifying the smallest units of human argumentation remains a key challenge for computational models of argument. This study tested the assumption that argumentative discourse units (ADUs) can be best described as clauses. Two online experiments investigated the role of ADUs in human language processing (Experiment 1) and recall (Experiment 2), providing evidence that discourse comprehension might be influenced by syntactic depth. Experiment 1 analysed participants' cued recall of pairs of clauses to identify whether they relied on clausal units when encoding information. Experiment 2 tested effects of manipulating the syntactic complexity (natural language – varying complexity, elementary language – one clause per sentence, or atomic language – one sub-clausal unit of information per sentence) on participants' free recall of short encyclopaedic entries adapted from *Wikipedia*. Both experiments found small-to-medium effects suggesting that defining ADUs as clauses might be justified to a degree, with potential implications for computational models of human argumentation.

Keywords. human argumentation, elementary discourse units, language comprehension, discourse processing, Rhetorical Structure Theory, memory recall

1. Introduction

Some of the key challenges of modelling argumentation computationally are the same as the challenges of understanding human language comprehension more generally. Theories of linguistic discourse describe human language processing beyond the sentence level through the existence of hierarchically organised relations between what Rhetorical Structure Theory (RST) has termed Elementary Discourse Units (EDUs) [1]. EDUs are commonly defined as elements of discourse which each contain one distinct proposition, i.e., a statement with a truth value [2], [3], linked through a variety of semantic relations [4], including ones comprising or giving rise to support or attack relations. To avoid potential pitfalls of presupposing the existence of elementary units of discourse processing, [5] proposed the concept of Argumentative Discourse Units (ADUs), referring to "minimal units of analysis" relevant for describing argumentation, although they note that ADUs might be equivalent to or contain multiple EDUs. Regardless of whether contexts of analysis might require EDUs or ADUs specifically [6],

[1] Corresponding Author: Clara Seyfried, c.seyfried@dundee.ac.uk.

the identification of minimal units might be considered the main goal of text segmentation in argument mining [7]. However, one key issue is that neither are definitively defined. As EDUs tend to be determined by functional rather than structural linguistic properties, their unit size is described as "arbitrary" [1]. Similarly, the Potsdam Commentary Corpus guidelines for RST segmentation [8] specify that the identification of ADUs should occur according to "units of meaning", which are said to frequently appear at the syntactic boundaries of main clauses. Yet despite theoretical considerations, there appears to be little empirical justification for how minimal units might be defined, and the practical consequences of different definitions remain poorly understood. Since they nevertheless continue to be used in models aiming to improve various aspects of argument mining, e.g., [9], [10], it is worth considering to what extent different definitions and usages of minimal units might currently converge.

Firstly, ADUs are often manually identified by human analysts, taking into account relevant social and linguistic context where appropriate. Human annotation of linguistic data can provide detailed analyses sensitive to context, yielding diverse training sets of natural language which can be used for machine learning, but it is also a slow, laboursome process which is highly dependent on the analysts' skillset and quality of instruction. So far, it remains unclear what determines how annotators intuitively segment text into ADUs, though variability has been noted to negatively affect inter-annotator agreement [11] [12]. Secondly, the size of ADUs might be defined linguistically, based on the rules of grammar. There has been a long tradition of defining EDUs as clauses, with occasional exceptions for sentential arguments and/or (restrictive) relative clauses [13]. Rhetorical Structure Theory describes EDUs as "essentially clauses", a definition [1] opted for because it is "theory-neutral", and [8] likewise define ADUs with reference to clauses. This approach can be applied easily but might fail to capture some of the subtleties of linguistic discourse. Thirdly, ADUs can also be determined computationally, e.g., [14]. [15] recently developed a segmentation model[2] that automatically splits text into EDUs using machine learning to identify yes/no boundary tags at the end of words. This is the quickest and easiest, but also the least fine-grained approach. Although [15] too describe EDUs as "clause-like", their model appears biased towards sub-clausal units. What these three current approaches have in common is that definitions of minimal units are not clearly justified, and so it remains uncertain whether similarities between the approaches and the results of their analyses are a sign of convergent validity or merely the result of a theoretical bias.

This study addresses this issue by investigating the potential role ADUs might play in human discourse comprehension, approaching the conceptual question of which linguistic units might be considered "elementary" through psycholinguistic evidence from the respective perspectives of human discourse processing (Experiment 1) and memory recall (Experiment 2). The main aim was to establish which level of syntactic depth can be best described as elementary, indirectly testing the usefulness of different definitions of ADUs by observing how humans represent linguistic information in memory. Based on some, albeit largely implicit agreement in the literature, it was hypothesised that clauses, rather than sub- (e.g., phrases) or super-clausal (e.g., sentences) units of linguistic information might determine humans' language processing and recall, as measured through their comprehension of written discourse.

[2] *SegBot:* http://138.197.118.157:8000/segbot/.

2. Related Work

So far, there is only limited empirical evidence that humans might process linguistic information in clauses, perhaps precisely because this is a fairly intuitive notion. When [3] first proposed that humans represent discourse in propositions, clauses might have been obvious examples because they are arguably the smallest linguistic units with a truth value (e.g., you can argue about whether "Dundee is beautiful", but not "Dundee is", or "beautiful", or even an individual morpheme such as "-ful"). Direct evidence for the propositional nature of clauses comes from a single study by [16], who tested priming effects in humans' single word recognition of sentences containing two clauses each and found that participants recognised words from previously studied sentences faster if they were preceded by words from the same, rather than a different clause from the same sentence. Since then, propositions have frequently been operationalised as clauses, e.g., [17], [18]. However, the usefulness of [3]'s propositional account has recently been challenged by [19], who tested it against an associative account of discourse comprehension. This account argues that discourse comprehension is not as much determined by logical relationships between units of information, but rather the closeness of humans' semantic representations, comparable to frameworks of graded semantics to model argumentation [20]. Acknowledging ensuing disagreement about the respective roles of grammar and pragmatics, this study therefore also took into account other linguistic factors, as well as cognitive and metacognitive abilities, including motivation.

Although little research so far has examined the comparative usefulness of different definitions of ADUs specifically, this study builds on substantial literature scattered between the discipline boundaries of philosophy, linguistics, and psychology. Propositional accounts of discourse processing follow the assumption that the ability to detect and understand causality (i.e., cause-effect relationships) is one of the key abilities of the human mind, see [21]. The bulk of research on humans' understanding of causality explains and partly justifies why theories of discourse such as Rhetorical Structure Theory built on the "causality-as-default" [22] assumption that humans automatically anticipate causal relationships between different units of text, see [23], although it is also possible that academia's logocentric legacies have led to a misleading over-emphasis of the importance of logical relationships in humans' communicative thought processes. Theoretical discussions of whether bipolar argumentation frameworks are adequate for describing argumentation [24] highlight the need to carefully evaluate the building blocks defining abstract and structured theoretical frameworks [25], [26]. For example, support and attack relations might drastically lose their meaningfulness depending on how similar or different the units they connect are. As [27] argue, formal theories of argumentation, such as abstract argumentation frameworks, can only be properly evaluated if they are empirically and theoretically validated. As it becomes evident that human argumentation is rife with fuzziness and ambiguity, which make it difficult to evaluate analyses of argumentation [12], theoretical approaches accounting for elements of unpredictability [28] might offer promising avenues for describing real-world human argumentation.

However, there is a lack of research attempting to link propositional and associative, logical and probabilistic, formal and informal models of argumentation. One notable exception appears to be [29], which systematically measured lexical, syntactic, structural, and pragmatic parameters of unit segmentation in machine learning and found structural and semantic features to be the most useful for identifying argument boundaries across

domains, although they note that the difference between EDUs and ADUs, as well as which factors influence unit boundaries remain important outstanding research questions. The present study attempts to inform future research by exploring these questions through a bottom-up approach, linking theoretical considerations to empirical observations of how humans remember linguistic information.

3. Experiment 1 (Discourse Processing)

3.1. Introduction

Language processing is frequently measured through eye-tracking, self-paced reading, or priming effects, such as [16], which found that words from previously studied sentences were recalled faster if they were preceded by words from the same, rather than a different clause from the same sentence. In order to test our main hypothesis, the assumption that ADUs can be best described as clauses, e.g., [30], we observed participants' ability to remember one but not both clauses in sentences consisting of pairs of clauses. Using cued recall as an indirect measure of discourse processing, we aimed to gain insights into the memory encoding process as observed through memory failures. At the same time, our choice of methodology was motivated by the relatively high real-world applicability of recall tasks, which we then built on further in Experiment 2.

The main analysis of this experiment compared the percentages of sentences for which each participant successfully remembered one, but not both of the two clauses present in each sentence with the percentages of sentences for which they had remembered parts of either clause, but neither in full. If the percentages of sentences for which participants remember exactly one clause are significantly higher, this could be taken as an indicator that participants might have encoded and retained the different clauses as individual units of information, and therefore provides support for the definition of ADUs as clauses.

3.2. Methods

3.2.1. Participants

A sample of 106 participants was recruited through social media. One participant was excluded because of a technical error. The remaining 105 participants were between 18 and 60 years old ($M = 28.41$, $SD = 9.66$), were proficient speakers of English (native or non-native), and did not have any perceptual, cognitive, or linguistic disorders. Most participants were native speakers of English ($N = 74$), but the sample also included native speakers of 16 other languages. The sample included 59 monolinguals, 26 bilinguals, and 20 multilinguals, representing over 21 nationalities.

3.2.2. Measures

Participants were asked to provide basic demographic information using a demographic questionnaire requiring them to indicate their age, gender, nationality, and level of education. This questionnaire also asked them to rate their general memory ability on a

6-point Likert scale. Participants' language background and reading behaviours were assessed using a short language proficiency questionnaire.

The main part of the experiment tested whether participants were more likely to correctly remember exactly one clause rather than only parts of each clause contained in sentences consisting of pairs of clauses. For this purpose, 18 sentences were constructed to be presented in a fixed order in three trials of six sentences each. Each sentence consisted of two clauses following a common syntactic structure (subject, verb, object), linked through the connective "and". Semantic connections between the clauses were intended to be plausible, but not strong, so as not to bias the results in either direction. The sentences described a variety of different topics and scenarios with both the gender (male, female, unspecified/gender-neutral) and degree of humanness (human, animal, neither, mixed) of agents balanced across trials. It was specified that sentences would be excluded from analyses if participants rated their plausibility at or below 2.5 on a 6-point Likert scale, but sentences reached an average rating of 4.57 ($SD = 0.53$), ratings ranging from 3.70 to 5.59. Recall was cued through reminders of one word per each of the six sentences, namely the subject, verb, or object of either clause of the respective sentences (Latin squares were used for partial counterbalancing, creating six lists). The order of cues was randomised. Just before recall, participants were asked to rate how much they thought they would be able to remember on a 6-point Likert scale (from "nothing" to "everything"). As there were exactly six sentences for each trial, this was used to measure participants' metacognition (i.e., they're ability to accurately self-evaluate).

3.2.3. Procedure

The experiment was conducted online using the experiment platform *Gorilla.*[3] Participants filled in the demographic and language questionnaires before completing three trials of the main task. For each trial, participants were reminded they would be shown six sentences for two minutes and instructed to read the information carefully so that they would be able to retain it as best as possible within that time limit. Participants were encouraged to read the information carefully multiple times, but instructed not to take any notes. A timer appeared after 1:30 minutes. After answering one metacognitive question, participants were asked to recall as much as they could remember from each set, prompted by one-word cues from each sentence. Finally, participants rated the plausibility of all 18 sentences on 6-point Likert scales and were given the option to describe their subjective memory experiences for each trial.

3.2.4. Analysis

Participants' recall for each sentence was scored twice and any disagreements were resolved. Recall was scored as correct if clauses and sentences were semantically correct, regardless of whether participants remembered sentences verbatim or paraphrased synonymous concepts. Responses were sorted into five different categories: sentences participants remembered completely (Category A), sentences for which participants remembered more than one clause but not both clauses correctly (Category B+), sentences for which participants remembered exactly one clause correctly (Category B), sentences for which participants remembered parts of both clauses but neither completely (Category C), sentences for which participants remembered less than a clause (Category D), and sentences participants did not recall at all (Category E). Recall cues were taken

[3] https://gorilla.sc/.

into account (e.g., responses correctly recalling a sentence's first clause prompted by a cue from the second clause were classified as B+, but responses only recalling the cue word were classified as E). The main analysis compared the number of participants' sentences in categories B and C using a paired-sample t-test.

3.3. Results

Table 1 shows participants' recall rates for each of the five main recall categories. Example responses illustrate that the categories were solely determined by correct semantic recall. For example, gendering a subject whose gender had not been specified resulted in this clause being counted as incorrect, i.e., equivalent to being incomplete. Note the high overall accuracy in the experiment. Two participants completed the task with 100% accuracy.

Table 1. Means and Standard Deviations of Participants' %s of Sentences in Each Category, with an Example Response for Each Category (Example Sentence: Trial 1, Sentence 1; Cues: N1: noun 1 ("baker"). V1: verb 1 ("opened"). O1: object 1 ("window"). N2: noun 2 ("priest"). V2: verb 2 ("spoke"). O2: object 1 ("prayer").)

Category (Correct)	*M*	*SD*	**"The baker opened a window and the priest spoke a prayer."**
A (both clauses)	52.34	2.48	"the baker opened the window and the priest said a prayer"^{O2}
B+ (more than one clause)	16.88	1.02	"The baker opened his window and the priest said a prayer."^{N2}
B (exactly one clause)	9.52	1.29	"The baker opened the window, and the"^{V1}
C (parts of both clauses)	4.02	0.51	"the monk said a prayer"^{N1}
D (less than one clause)	3.17	0.59	"... opened the window and ..."^{O1}
E (parts of neither clause)	14.18	1.43	"The opened and"^{V1}

A paired-sample t-test compared participants' mean percentages of responses in categories B and C and found that responses of type B were statistically significantly more common than responses of type C, $t(104) = 3.82$, $p < .001$, two-tailed; in fact, more than twice as common on average. Post-hoc power analyses were conducted in G*Power [31]. Based on the small-to-medium effect size of the difference between the two categories (Cohen's $dz = 0.37$) at $\alpha = .05$ with a sample of $N = 105$, the analysis appears to have detected the effect with 95% power. It should be noted that none of the distributions for the different response category ratios was normally distributed (Shapiro-Wilk, $W(105) = .57$ to 0.98, $p = 0.048$ to $p < .001$), but despite the data being left-skewed, it is expected that the t-test was robust and statistically powerful enough so as not to be affected by the violation of the normality assumption. Since participants remembered the meaning of individual clauses together more frequently than they only remembered parts of different clauses, this analysis suggests evidence that participants might have processed the sentences as clauses. Similarly, out of the total of 319 sentences in Category B+, 305 (i.e., 95.61%) correctly described the clause containing the cue word, rather than participants having been prompted by the cue word of one clause to remember the other clause in its entirety.

3.4. Discussion

Since participants were more likely to remember exactly one clause of each sentence rather than only parts of each clause, this experiment provides some evidence supporting the hypothesis that ADUs might be described as clauses in discourse processing, as measured indirectly through cued recall. The nature of the stimuli used, with their symmetrical structure, separated by the conjunction "and", might have biased participants to encode the information in clausal units. Nevertheless, stimuli were deliberately designed to be interpretable as either two separate or one individual event, and participants' plausibility ratings confirmed this to be the case. The small-to-medium effect therefore provides some evidence indicating that the definition of ADUs as clauses holds at least in fairly regular linguistic contexts. However, it is less clear to what extent the observed effect might be observed under more naturalistic conditions, which might be crucial for informing how ADUs should be determined to model argumentation.

4. Experiment 2 (Memory Recall)

4.1. Introduction

To extend the findings of Experiment 1, the second experiment focused less on assessing the function of ADUs during language processing, but instead determined the real-world applicability of the concept directly by considering whether potential effects of manipulating syntactic depth would be observable under ecologically valid conditions. More specifically, this experiment tested whether manipulating the syntactic complexity (natural language – varying complexity, elementary language – one clause per sentence, or atomic language – one sub-clausal unit of information per sentence) would influence participants' memory recall of information contained in short encyclopaedic entries adapted from *Wikipedia* [4]. Initially, this experiment also manipulated the texts' presentation format (paragraph or bullet points), but as power analyses suggested that the experiment had been severely underpowered to determine effects of the size observed in Experiment 1, these analyses are based on a sub-sample of the original experiment.

The assumption that sentences function as discrete units of information during language processing is even more widespread than that clauses represent propositions, [3], [32]. Indeed, both are closely related, as there is no difference between sentence and clausal units in simple sentences, which, by definition, only consist of one (independent) clause. This tends to be reflected in guidelines for annotators of human argumentation, e.g., [8], [30]. By adapting the syntactic complexity of natural language for the elementary and atomic language conditions (the natural language condition representing an experimental control), this experiment not only explored whether the use of sub-, or clausal units of information would improve memory recall compared to naturally occurring language, but also thereby simultaneously tested a method for simplifying texts to facilitate explainability of human argumentation, should this be the case.

Assuming that clauses function as ADUs, it was hypothesised that participants exposed to elementary language would recall information better than participants in the natural and atomic language conditions.

[4] https://www.wikipedia.org/.

4.2. Methods

4.2.1. Participants

A sample of 59 participants was recruited through social media. Following the exclusion of two outliers, whose performance changed by more than 2.5 standard deviations from the participant mean from baseline, the final sample consisted of 57 participants. All participants were between 18 and 55 years old (M = 28.26, SD = 8.86), were proficient speakers of English (native or non-native), and did not have any perceptual, cognitive, or linguistic disorders. Most participants were native speakers of English (N = 36), but the sample also included native speakers of twelve other languages, including 30 monolinguals, 19 bilinguals, and eight multilinguals, representing over 15 nationalities.

4.2.2. Measures

Participants were asked to provide information about their basic demographic background, general memory ability, and language background using the same demographic and language questionnaires as in Experiment 1.

The main part of the experiment manipulated the syntactic complexity of three short introductions to encyclopaedic entries adapted from Wikipedia. All three were texts about mythological creatures, presented in a fixed order (baseline: dragons, trial 1: unicorns, trial 2: phoenixes). For the baseline trial and the control condition, the texts were slightly shortened (and somewhat adapted) versions of the original entries from Wikipedia (natural language: NL). These texts were of comparable length for the baseline (127 words in six sentences) and the two experimental trials (129 words in six sentences and 127 words in six sentences). To manipulate syntactic complexity, the NL texts were adapted for conditions containing either one clause per sentence (elementary language: EL) or one sub-clausal unit of information per sentence (atomic language: AL). For the EL condition, sub-ordinate clauses were transformed into new sentences, whereas the AL condition also separated other dependent clauses containing unique units of information, classifying adjectives, and conjuncts coordinated by “and”. For example, consider the NL text for trial 2: “The phoenix is an immortal bird associated with Greek mythology (with analogues in many cultures) that cyclically regenerates or is otherwise born again”. For the EL condition, this was changed to “The phoenix is an immortal bird associated with Greek mythology (with analogues in many cultures)” and “The phoenix cyclically regenerates or is otherwise born again”. For the AL condition, these sentences were segmented further, resulting in: “The phoenix is immortal”, “The phoenix is a bird”, “The phoenix is associated with Greek mythology”, and “The phoenix has analogues in many cultures”. As the segmentation of sentences required the reconstruction of anaphora, the text length increased to 134-136 words in 8-10 sentences (EL) and 169-180 words in 11-19 sentences (AL) for the experimental trials.

In addition to participants’ ratings of their subjective memory ability in the demographic questionnaire, a range of metacognitive factors were considered for each trial. Specifically, participants rated their difficulty reading, their difficulty understanding, personal interest in, and prior knowledge of the topic covered in each text on 6-point Likert scales. Participants also indicated how much they thought they would be able to remember from the information they had read.

4.2.3. Procedure

Like Experiment 1, Experiment 2 was conducted online using *Gorilla*. Participants filled in the demographic and language questionnaires before completing three trials of the main task. For the baseline trial, participants were shown a short text for two minutes (participants received the same instruction as in Experiment 1). A timer appeared after 1:30 minutes. After answering five metacognitive questions about the text, participants were asked to freely recall as much as they could remember from the information they had read. Then, each participant was randomly allocated to one of the experimental conditions and the same procedure was repeated for the two experimental trials.

4.2.4. Analysis

Participants' free recall was manually scored and calculated as percentages of marks (or half-marks) gained according to predefined criteria. Participants could gain marks for remembering specific words or concepts. No marks were deducted if participants remembered information that was incorrect. The main analysis compared mean changes in recall rate from baseline using one-way between-subjects analysis of covariance (ANCOVA) with three levels of syntactic complexity (NL, EL, AL) and performance at baseline and topic knowledge as covariates reflecting general memory ability. Additional analyses used similar analyses of variance (ANOVAs) with metacognitive variables as the dependent variable.

4.3. Results

Table 2. Means and Standard Deviations of Participants' Free Memory Recall (in %) for Three Conditions (NL, EL, AL) and Three Texts (Baseline: Dragons, Trial 1: Unicorns, Trial 2: Phoenixes), Changes in Recall Rates from Baseline to the Mean of Trials 1-2 (in % points), and Average Topic Knowledge Ratings (out of 6).

Condition		Baseline	Trial 1	Trial 2	Mean Change	Topic Knowledge
Natural Language (N = 13)	*M*	64.5	67.3	60.0	-0.006	2.62
	SD	10.4	11.4	16.6	0.081	1.43
Elementary Language (N = 12)	*M*	62.9	73.1	59.6	0.034	2.79
	SD	7.2	5.8	10.0	0.097	0.84
Atomic Language (N = 32)	*M*	63.3	67.5	52.5	-0.034	2.88
	SD	12.8	13.5	15.0	0.085	1.23

Table 2 shows the descriptive statistics of participants' recall in the three conditions for the baseline and experimental trials, as well as the changes in recall rates from baseline and average topic knowledge ratings per participant.

A one-way between-subjects ANCOVA tested whether syntactic complexity had affected participants' changes in memory recall, controlling for baseline performance (indexing general memory ability) and topic knowledge (the only metacognitive factor correlated with recall in the wider experiment, $N = 101$, $M = 3.67$, $SD = 1.23$; $r = .29$, $p < .01$). The main prediction was that participants exposed to EL would recall more information than participants in the NL and AL conditions. This analysis found a statistically significant effect of syntactic complexity on change in recall performance, $F(2, 52) = 3.23$, $p = .0.047$, $\eta_p^2 = .09$. Participants who received EL texts were the only condition to improve on average, with performance in the NL and AL conditions decreasing from baseline. A post-hoc Tukey HSL test ($\alpha = 0.05$) showed that the

differences between change in recall rate was statistically higher for the EL as opposed to the AL condition, whereas there were no differences between NL and EL.

Further 3 x 2 between-subjects ANOVAs tested whether the main manipulations had influenced a range of metacognitive factors. Indeed, there was a main effect of syntactic complexity on participants' changes in perceptions of the texts' linguistic complexity, $F(2, 54) = 4.30$, $p < .02$, $\eta_p^2 = .13$. A post-hoc Tukey HSL test ($\alpha = 0.05$) showed that AL was perceived as more difficult to read than NL (M =1.31, SD = 1.15, compared to M = 0.42, SD = 0.76), though not the EL condition (M = 0.67, SD = 0.81). Similarly, there was an effect of syntactic complexity on participants' changes in interest in the texts, $F(2, 54) = 3.49$, $p < .04$, $\eta_p^2 = .11$, although a post-hoc Tukey HSL test found no significant differences (M = -0.31, SD = 0.83 for NL, M = -0.50, SD = 0.80 for EL, and M = -1.09, SD = 1.53 for the AL condition). There were no statistically significant effects of syntactic complexity on participants' ratings of their difficulty understanding, prior knowledge of, or predictions about remembering information from the texts.

4.4. Discussion

As a medium-sized statistically significant effect of syntactic complexity on changes in recall of written discourse could be observed, Experiment 2 found some evidence that defining ADUs as clauses might help model human memory recall of linguistic discourse. However, it should be noted that although participants' performance improved most for participants in the EL condition on average, this condition was not significantly different from the NL condition. Participants in the AL condition, who received texts adapted so that individual points of information would be contained in what were sub-clausal units of information in the original texts, found those texts to be more linguistically complex, might have lost more interest in, and performed worse at remembering the texts compared to participants in the other two conditions. An explanation for this appears to be that there were only very slight differences between the texts for the NL and EL conditions, as many sentences in natural language already only contain one clause. While the use of highly naturalistic stimuli is a clear advantage of this experiment, the use of only three trials in total meant that the diversity of linguistic phenomena contained in the texts, including ones requiring judgments about what might be interpreted as a clause, was very limited. The experiment would therefore benefit from a replication with a greater sample size and/or more trials.

Generally, the results of this experiment suggest that although there seems to be some evidence that defining ADUs as clauses might be justified. However, future research should consider distributions of clausal-sized units in natural language, further investigating what determines how humans segment text into units of potentially different sizes, ideally by comparing NL and EL conditions of argumentative texts. That slight changes in the syntactic structure of texts were associated with measurable changes in participants' perceived complexity and actual recall of the texts further highlights a potential for future research into improving the explainability of argumentation [33].

5. Discussion and Conclusion

This study investigated the role of ADUs in discourse comprehension by testing whether information is processed and remembered in clausal as opposed to larger or smaller syntactic units. In Experiment 1, participants were more likely to remember exactly one

rather than only parts of either clause contained in sentences consisting of pairs of clauses. Experiment 2 found that the syntactic complexity of short encyclopaedic texts influenced participants' free memory recall of the information they contained, with sub-clausal unit sizes impairing participants' memory performance, indicating that it might be justified to define ADUs as clauses to a degree.

These findings are much in line with the predictions of Rhetorical Structure Theory [1], which describes discourse through semantic relations between clausal ADUs. The results of Experiment 1 complement [16]'s finding of clausal priming effects in sentence recognition by measuring language processing indirectly through cued recall. Indeed, one of the key strengths of this study is that it tested effects of syntax on discourse comprehension using two different, but closely related approaches, and, although the main results of Experiment 2 did not find a significant difference between natural and elementary language, nevertheless found comparable numerical differences. Similarly, effects emerged despite the linguistically diverse participant sample, suggesting they might be relatively robust. However, Experiment 2 in particular would benefit from a (conceptual) replication. Evidence that manipulating syntactic complexity might facilitate human recall of linguistic information could have wide-ranging practical and theoretical implications for the study of human argumentation and beyond.

Although this study has found some empirical evidence that justifies basing discourse analyses on the definition of ADUs as clauses, it is unlikely that syntax alone can explain how humans segment discourse. As [19] point out, past research might have neglected the importance of lexical associations between units [34], and semantics appear to be highly important within specific domains [29]. However, abstract argumentation frameworks might model the complexities of human argumentation more effectively if propositions are adequately defined [24]. Determining ADU unit sizes might be as crucial as identifying argument relations, as the effectiveness of modelling argumentation is dependent on the relevance of relations for the arguments they aim to represent. As [35] show, argument relation classification might be more robust if systems are trained on discourse context and argument content separately. At the same time, a better understanding of ADUs can also inform research on discourse relations, which can in turn validate and thereby help unify theories of both discourse generally and argumentation specifically. Ultimately, though further research is needed, understanding the role of ADUs might therefore be a suitable starting point for future argument mining models that describe not only how arguments could evolve, but also how humans argue.

References

[1] Mann WC, Thompson SA. Rhetorical structure theory: toward a functional theory of text organization. Text - Interdisciplinary Journal for the Study of Discourse. 1988;8(3):243-81.
[2] Walton DN. What is reasoning? What is an argument? Journal of Philosophy. 1990;87(8):399-419.
[3] Kintsch W, Dijk TAv. Toward a model of text comprehension and production. Psychological Review. 1978;85(5):363.
[4] Taboada M, Mann WC. Rhetorical structure theory: looking back and moving ahead. Discourse Studies. 2006;8(3):423-259.
[5] Stede M, Afantenos S, Peldszus A, Asher N, Perret J. Parallel discourse annotations on a corpus of short texts. 10th International Conference on Language Resources and Evaluation (LREC 2016); 2016. p. 1051-1058.
[6] Peldszus A, Stede M. From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence. 2013;7(1):1-33.
[7] Lawrence J, Reed C. Argument mining: A survey. Computational Linguistics. 2019;45(4):765-818.

[8] Stede M, Mamprin S, Peldszus A, Herzog A, Kaupat D, Chiarcos C, et al. Handbuch Textannotation Potsdamer Kommentarkorpus 2.0. Stede M, editor. Potsdsam: Universitätsverlag Potsdam; 2016. 234 p.
[9] Saha S, Das S, Srihari R, editors. EDU-AP: Elementary Discourse Unit based Argument Parser. Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue; 2022; Edinburgh, UK: Association for Computational Linguistics.
[10] Xiong Y, Racharak T, Nguyen ML, editors. Extractive Elementary Discourse Units for Improving Abstractive Summarization. SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information; 2022; Madrid, Spain.
[11] Hasan KS, Ng V, editors. Why are you taking this stance? Identifying and classifying reasons in ideological debates. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
[12] Hautli-Janisz A, Schad E, Reed C, editors. Disagreement space in argument analysis. 1st Workshop on Perspectivist Approaches to NLP (LREC2022); 2022; Marseille, France: European Language Resources Association (ELRA).
[13] Schauer H, editor From elementary discourse units to complex ones. SIGDIAL '00: Proceedings of the 1st SIGdial workshop on Discourse and dialogue; 2000; Stroudsburg, PA,.
[14] Prevot L, Hunter J, Muller P, editors. Comparing Methods for Segmenting Elementary Discourse Units in a French Conversational Corpus. 24th Nordic Conference on Computational Linguistics (NoDaLiDa 2023); 2023; Tórshavn, Faroe Islands, Finland. .
[15] Li J, Sun A, Joty S. SegBot: A Generic Neural Text Segmentation Model with Pointer Network. International Joint Conference on Artificial Intelligence; 2018; Stockholmsmässan, Sweden.
[16] Ratcliff R, McKoon G. Priming in item recognition: Evidence for the propositional structure of sentences. Journal of Verbal Learning and Verbal Behavior. 1978;17(4):403-17.
[17] Corro LD, Gemulla R, editors. ClausIE: clause-based open information extraction. WWW '13: Proceedings of the 22nd international conference on World Wide Web; 2013; Rio de Janeiro, Brazil.
[18] Park J, Cardie C, editors. Identifying Appropriate Support for Propositions in Online User Comments. Proceedings of the first workshop on argumentation mining; 2014; Baltimore, Maryland.
[19] Shabahang K, Yim H, Dennis SJ. Associations versus propositions in memory for sentences. CogSci. 2019.
[20] Thimm M, Cerutti F, Rienstra T. Probabilistic Graded Semantics. Computational Models of Argument: IOS Press; 2018. p. 369-80.
[21] Diessel H, Hetterle K. Causal clauses: a cross-linguistic investigation of their structure, meaning, and use. Linguistic Universals and Language Variation: de Gruyter; 2011. p. 21-52.
[22] Sanders TJM, editor Coherence, Causality and Cognitive Complexity in Discourse. Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning; 2005.
[23] Schölkopf B. Causality for Machine Learning. Probabilistic and Causal Inference: The Works of Judea Pearl: ACM Books; 2022. p. 765–804.
[24] Amgoud L, Cayrol C, Lagasquie-Schiex MC, Livet P. On bipolarity in argumentation frameworks. International Journal of Intelligent Systems. 2008;23(10):1062-93.
[25] Dung PM. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence. 1995;77(2):321-57.
[26] Prakken H. An abstract framework for argumentation with structured arguments. Argument & Computation. 2010;1(2):93-124.
[27] Prakken H. Argumentation Frameworks: the Case of Bipolar Argumentation Frameworks. Computational Models of Natural Argument. 2020(2669):21-30.
[28] Hahn U. Argument quality in real world argumentation. Trends in Cognitive Sciences 2020;24(5):363-374.
[29] Ajjour Y, Chen W-F, Kiesel J, Wachsmuth H, Stein B, editors. Unit segmentation of argumentative texts. Proceedings of the 4th Workshop on Argument Mining; 2017; Copenhagen, Denmark: Association for Computational Linguistics.
[30] Ruiz-Dolz R, Nofre M, Taulé M, Heras S, García-Fornes A. VivesDebate: A New Annotated Multilingual Corpus of Argumentation in a Debate Tournament Applied Sciences. 2021;11(15).
[31] Faul FE, Lang A-GE, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods. 2007;39(2):175–91.
[32] King J. Structured propostions and sentence structure. Journal of Philosophical Logic. 1996;25:495-521.
[33] Prakken H, Winter Md. Abstraction in Argumentation: Necessary but Dangerous. Computational Models of Argument: Proceedings of COMMA 2018. p. 85-96.
[34] Carstens L, Toni F, editors. Towards relation based argumentation mining. Proceedings of the 2nd Workshop on Argumentation Mining; 2015; Denver, Colarado (USA).
[35] Opitz J, Frank A. Dissecting Content and Context in Argumentative Relation Analysis. Proceedings of the Sixth Workshop on Argument Mining; 2019; Florence, Italy.