Analyzing Culture-Specific Argument Structures in Learner Essays

Wei-Fan Chen
Paderborn University
Department of Computer Science
cwf@mail.upb.de

Mei-Hua Chen
Tunghai University
Department of Foreign Languages and Literature
mhchen@thu.edu.tw

Garima Mudgal
Paderborn University
Department of Computer Science
garima@mail.upb.de

Henning Wachsmuth
Paderborn University
Department of Computer Science
henningw@upb.de

Abstract

Language education has been shown to benefit from computational argumentation, for example, from methods that assess quality dimensions of language learners' argumentative essays, such as their organization and argument strength. So far, however, little attention has been paid to cultural differences in learners' argument structures originating from different origins and language capabilities. This paper extends prior studies of learner argumentation by analyzing differences in the argument structure of essays from culturally diverse learners. Based on the ICLE corpus containing essays written by English learners of 16 different mother tongues, we train natural language processing models to mine argumentative discourse units (ADUs) as well as to assess the essays' quality in terms of organization and argument strength. The extracted ADUs and the predicted quality scores enable us to look into the similarities and differences of essay argumentation across different English learners. In particular, we analyze the ADUs from learners with different mother tongues, different levels of arguing proficiency, and different context cultures.

1 Introduction

Analyzing the argument structure of a text helps understand the individual points being made and the relationships between these points to identify the overall position that the writer supports (Lawrence and Reed, 2020). In practice, manual annotation of argument structure is a skilled work; the laborious and time-consuming process behind would make large-scale studies challenging. This is undoubtedly true for second-language writing research. Especially studies investigating language learners' use of arguments in the essays usually need to determine the occurrence of individual argument components, such as Paek and Kang (2017) and Liu and Wan (2020), see Section §2 for details.

Research on computational argumentation has drawn increased interest in recent years, with argumentative writing support being one of the main envisioned applications (Stab and Gurevych, 2017; Wambsganss and Niklaus, 2022). Computational methods to automatically mine argumentative discourse units (ADUs) and the relations between these units enable various applications in the context of language education (Wambsganss et al., 2021; Putra et al., 2021). Argument mining has been performed effectively on persuasive learner essays (Stab and Gurevych, 2014b), and argument quality assessment has been aided with claim generation (Gurcke et al., 2021). Given the close connection between argument structure and text quality (Putra et al., 2021), argumentative learner essays have also been studied in terms of quality dimensions such as organization (Persing et al., 2010) and argument strength (Persing and Ng, 2015).

So far, however, little attention has been paid to the cultural diversity of language learners with respect to the different argument structures they form. Cultural variation may originate from different geographical origins, mother tongues, societal systems, the ways people communicate in these systems, and many other aspects (Senthamarai and Chandran, 2015). Some of these aspects may be easy to access, others barely. Either way, culture is recognized known as a factor affecting the persuasiveness of arguments and the organization of ideas of language learners (Carlile et al., 2018; Putra et al., 2021). At the same time, the extent to which culture is reflected in a given text may depend on the learner's level of language proficiency. Bearing these points in mind, this paper goes beyond previous studies of learner argumentation, analyzing differences in the structure and quality of essay argumentation of culturally diverse learners.[1]

[1] Studying cultural differences in the context of text quality

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 51–61
October 17, 2022.

To learn about cultural differences, we first build statistical and neural NLP models, following previous research, to classify ADUs in learner essays and to extract common structural argument patterns in terms of sequences of types of ADUs in a paragraph (hereafter, *ADU flows*). Moreover, in line with the a study of the impact of argument structure on text quality (Wachsmuth et al., 2016), we develop models to score the essays in terms of their organization and argument strength.

First, a state-of-the-art approach (Prakash and Madabushi, 2020) is adapted for mining ADUs from English texts, trained on the 402 persuasive student essays of Stab and Gurevych (2017) as well as on a corpus of Reddit ChangeMyView discussions (Hidey et al., 2017). Then, two scoring models are learned on 1000 essays from the ICLE corpus (Granger et al., 2009; Persing et al., 2010; Persing and Ng, 2015, Section §3). The trained models are compared with two strong baselines, including the current state of the art on the the respective tasks (Section §4), in order to get an idea of their reliability. The models then serve as the basis for the main analysis carried out in this paper.

In particular, applying the trained models to the entire ICLE corpus, we contrast the most frequent ADU flows that learners use depending on their cultural background, reflected in the author's first language, and in terms of whether that is a high or low-context language (Hall, 1976), as well as their proficiency levels, reflected in the scores of organization and argument strength (Section §5). In addition, we analyze the macro-structure of the essays to classify essays into climactic/anti-climactic (Suzuki, 2010) and horizontal/vertical (Suzuki, 2011). The results suggested that the most frequent ADU flows and their macro-structures correlate with the cultural background and language proficiency of learners, revealing various patterns. For example, speakers of European languages tend to use similar ADUs flows, and among them, speakers of Germanic language have even more similar ADUs flows.

Altogether, we make three contributions in this analysis-oriented paper:

1. We present computational methods that reliably mine ADUs from persuasive essays and that score the essays' quality.
2. We extend computational research on essay argumentation by the consideration of cultural differences between the essays' authors.
3. We provide meaningful insights into the similarities and differences of essay argumentation across different English learners.

The code of our experiments is available at: https://github.com/webis-de/argmining22-culture-arg.

2 Related Work

Most research on language learners' argumentation competence investigates essays of a small number of ESL learners in their own countries, such as Paek and Kang (2017) and Liu and Wan (2020). Paek and Kang (2017) study how Korean students use Toulmin elements in their English essays. The results show that Korean students relied heavily on claim and data due to the Korean culture-specific discourse. Liu et al. (2019) and Qin and Karabacak (2010) analyze Toulmin elements in Chinese students in their English argumentative writings. The researchers find that Chinese students mainly use data and subclaim but they barely use counterarguments and rebuttal to consider opposing views. In addition, influenced by Chinese culture, Taiwanese students prefer backing and modal besides data and claim (Cheng and Chen, 2009). The study of Abdollahzadeh et al. (2017) on Iranian graduate learners of English also shows that the students are prone to use data and claim the most.

On the other hand, numerous studies (Kim, 1997; Suzuki, 2010; Kim et al., 2011; Suzuki, 2011; Liu and Furneaux, 2014; Vajjala, 2018) have investigated the effects of culture on persuasive essays produced by native and non-native learners. For example, Kim (1997) studies the differences in Korean and American editorials while Suzuki (2010) conducts a similar study that compares the arguments written by Japanese and American. The results show that non-native students tend to transfer their first language rhetorical style into their English writing. Particularly, non-native speakers tend to use climactic and vertical macro-structures while English speakers tend to use anti-climactic and horizontal macro-structures. These terms are elaborated in Section §5.1.

The above mentioned studies suggest the learners' argument structures would differ depending on their mother tongue backgrounds. Language is the carrier of culture, and cultural features can be

is, by concept. an ethically sensitive endeavor. We point out already here that we do not assess whether people from some cultures argue "better" than others, but to learn about differences in arguing that may be important to provide adequate writing support (see Section §8 for details).

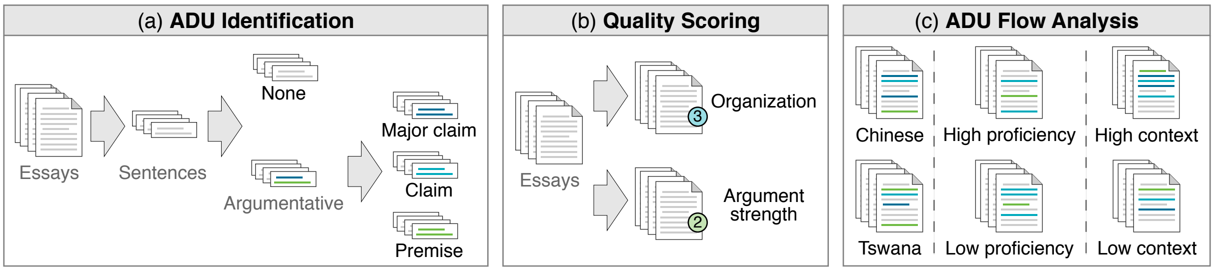


Figure 1: Overview of this paper: (a) We identify sentence-level argumentative discourse units (ADUs) in essays distinguishing four types: *none*, *major claim*, *claim*, and *premise* (Section §3.1). (b) We score the essays' quality in terms of *organization* and *argument strength* (Section §3.2). (c) We analyze ADU flows of cultural diverse learners in terms of first language (Section §5.1), arguing proficiency (Section §5.2) and language context (Section §5.3).

reflected in one's writing. Hall (1976) suggests the categorization of cultures into high context versus low context cultures[2] in order to understand their basic differences in communication style and cultural issues. In fact, the communication styles of people from different cultures range from explicit to ambiguous (Hall, 1976; Zou, 2019; Panina and Kroumova, 2015). That means one culture is more or less high-context (or low-context) than the other. Zou (2019) shows various cultures on a continuum, from where a tendency is observed that most Northern European countries are low-context whereas Asian countries are more high-context. Similarly, Senthamarai and Chandran (2015) classifies North America and much of Western Europe are low-context while Middle East, Asia, Africa, and South America are high-context. Given that the "thought patterns" (Kaplan, 1966) are expected as an integral part of their communication, the "cultural thought patterns" (Kaplan, 1966) may affect the persuasiveness of arguments and organization of ideas (Carlile et al., 2018; Putra et al., 2021).

With a better understanding of how learners from different cultural groups write arguments, language teachers could help learners enhance the quality of argumentative writing. Unfortunately, the impact of cultural differences on argument forms of learners from diverse mother tongue backgrounds is understudied. Typically, such studies rely heavily on manual annotation of argument structures. It is a skilled work. The laborious and time-consuming process would make large-scale studies challenging. Luckily this thorny issue could be addressed using argument mining technology. It has enabled a variety of applications (Wambsganss et al., 2021).

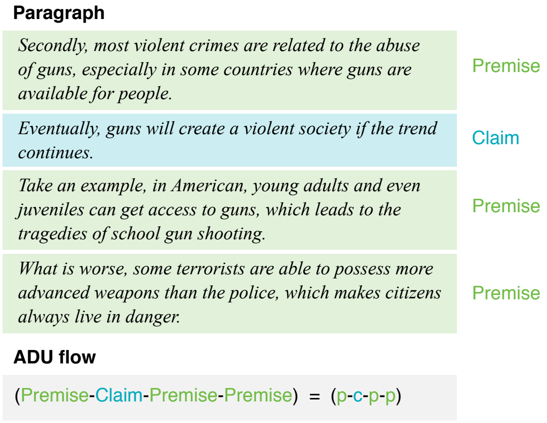
To achieve our goal, two main tasks are performed. Mining argumentative discourse units (ADUs) is the first task of most argumentation technologies. The argument annotated essays corpus (Stab and Gurevych, 2014a), has been widely used to find the boundaries of ADUs with sequential labeling (Stab and Gurevych, 2017; Ajjour et al., 2017), to identify the types of ADUs (Stab and Gurevych, 2014b), or recognize the relations between ADUs (Stab and Gurevych, 2014b). A subsequent computational argumentation task is to assess the essay quality. The International Corpus of Learner English (Granger et al., 2009) has been adopted to assess various quality dimensions of persuasive essays, such as organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014) and argument strength (Persing and Ng, 2015). In this paper, we target the two fundamentally important dimensions: organization and argument strength.

We build on the setting of Wachsmuth et al. (2016), but analyze a large number of different learner populations. To the best of our knowledge, we are the very first paper aiming at providing an in-depth analysis of ADUs produced by learners from various cultures, and revealing the differences and similarities of argument structures among different learner populations and proficiency levels from the perspective of computational argumentation.

3 Method

This section presents the computational methods that we develop for identifying argumentative discourse units (ADUs) in student essays as well as for scoring quality dimensions of the essays. We discuss how the methods are trained and what features are used. Figure 1a and b illustrate their usage.

[2] High and low-context will be discussed in Section§ 5.3.



Paragraph

Secondly, most violent crimes are related to the abuse of guns, especially in some countries where guns are available for people. Premise

Eventually, guns will create a violent society if the trend continues. Claim

Take an example, in American, young adults and even juveniles can get access to guns, which leads to the tragedies of school gun shooting. Premise

What is worse, some terrorists are able to possess more advanced weapons than the police, which makes citizens always live in danger. Premise

ADU flow

(Premise-Claim-Premise-Premise) = (p-c-p-p)

Figure 2: Argumentative discourse units (ADUs) and an ADU flow. The example is adapted from Wachsmuth et al. (2016). This paragraph contains three premises and one claim in the order of *premise-claim-premise-premise*.

3.1 ADU Identification

In this study, we see ADU identification as classifying each sentence of an essay into one of four types: *major claim*, *claim*, *premise*, and *none*. In line with Stab and Gurevych (2014b), we decompose the task into two stages, as in Figure 1a: the first separates all sentences into non-argumentative units (*none*) and argumentative units. In the second stage, another model classifies each argumentative unit into *major claim*, *claim*, and *premise*. Inspired by Prakash and Madabushi (2020), we use multilayer perceptron (MLP) in both stages whose features are TF-IDF values of words and the sentence embedding vector encoded by RoBERTa (Liu et al., 2019).

After extracting the ADUs in an essay, we then identify the ADU flows as the ADU type sequence in a paragraph. As shown in Figure 2, given that there are ordered *premise, claim, premise, and premise* in the paragraph, the ADU flow here is *premise-claim-premise-premise*, or *p-c-p-p* for short.

3.2 Quality Scoring

As shown in Figure 1b, we use two scoring models to predict the quality of essays on a 4-point scale, in terms of their organization and argument strength, respectively. For scoring, we employ random forest regression (Breiman, 2001). The models' features combine distributed semantics with structure-oriented features handcrafted for the given task. In particular, for distributed semantics, we make use of the last hidden layer of BERT. Conceptually, this layer should encode the meaning of the input in the form of a vector. For the handcrafted features, we reimplement a set of features mostly proposed by Wachsmuth et al. (2016), namely:

ADU type	Training	Validation	Test	Total
Major Claim	520	93	80	693
Claim	1,698	306	190	2,194
- Claims (AAE)	1,016	183	190	1,389
- Claims (CMV)	682	123	0	805
Premise	2,515	441	450	3,406
None	997	172	168	1,337

Table 1: The number of ADU types in the training, validation, and test sets built from the employed corpora.

- Frequencies of nouns, verbs, and adjectives in the essay
- ADU n-grams in the essay, with $n \in \{1, 2, 3\}$
- ADU compositions, i.e., frequencies of combinations of ADU types within paragraphs
- ADU flows, i.e., sequences of ADU types (or changes thereof) within paragraphs
- Paragraph flows, i.e., sequences of discourse functions: introduction, body, and conclusion (Persing et al., 2010)

3.3 Data and Experiments

For ADU identification, we employ the Argument Annotated Essays (AAE) corpus of Stab and Gurevych (2017). As the number of claims is rather small in the corpus, we include claims from the ChangeMyView (CMV) corpus annotated by Hidey et al. (2017).[3] Following Wachsmuth et al. (2016), we treat ADU identification as a sentence-level classification task: A sentence is labeled with one of the classes if any part of the sentence is labeled with that class. After merging the two corpora, we randomly split them into training, validation, and test sets by a 70-15-15 split. The distribution of the ADU types in the datasets can be seen in Table 1.

As for the quality scoring task, we rely on the annotated subset of 1000 essays from the ICLE corpus (Persing et al., 2010; Persing and Ng, 2015). We use the same splitting and the 5-fold setting as Wachsmuth et al. (2016). The distribution of the scores can be seen in Table 2.

[3] We use the authors' updated corpus version: `https://github.com/chridey/change-my-view-modes`. Thus, the data distribution differs from Hidey et al. (2017).

Quality Dimension	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Organization	24	14	35	146	416	289	79
Argument strength	2	21	116	342	372	132	15

Table 2: The number of essays of each score for argument strength and organization in the data employed from Persing et al. (2010) and Persing and Ng (2015).

ADU Identification	M.Cl.	Claim	Prem.	Macro
Baseline majority	0	0	67.3	22.4
Baseline SVM	54.6	23.5	70.4	49.5
Our method w/o CMV	77.6	57.5	83.9	73.0
Our method	85.0	67.8	88.5	80.4
Stab and Gurevych (2017)	89.1	68.2	90.3	82.6

Table 3: Effectiveness of two variations of our ADU identification method and the baselines. The columns show the F_1-score for major claims (M.Cl.), claims, premise (Prem.), and macro F_1-score.

4 Results

We seek to apply ADU identification and quality scoring in order to analyze the whole ICLE corpus with 6,085 essays in total. This section discusses the effectiveness of the trained models.

4.1 ADU Identification

We compare our method to two baselines, a majority baseline and an SVM based on word 1-, 2- and 3-grams, as well as to Stab and Gurevych (2017). As seen in Table 3, our approach outperforms both baselines with a large margin. It also shows that adding claims from CMV improves the performance in all regards. Compared to Stab and Gurevych (2017), our method does not perform better mainly because of the limited comparability. Our evaluation is performed at the sentence level whereas theirs is a token-based evaluation.

4.2 Quality Scoring

The effectiveness of our scoring models are compared to the results of Persing et al. (2010), Persing and Ng (2015), and Wachsmuth et al. (2016) in Table 4. With respect to argument strength and organization, our method performs better than Persing et al. (2010) and Persing and Ng (2015) but worse than Wachsmuth et al. (2016) in terms of MAE and MSE. Our organization scoring model is almost on par with the others. The difference could result from the features; we employ BERT encodings while Wachsmuth et al. (2016) fine-tuned handcrafted semantic features.

	Arg. Strength		Organization	
Approach	MAE	MSE	MAE	MSE
Persing et al. best	0.392	0.244	0.323	0.175
Wachsmuth et al. best	0.378	0.226	0.314	0.167
Our approach	0.385	0.229	0.346	0.193

Table 4: Effectiveness of our quality scoring methods compared to previous approaches in terms of mean absolute error (MAE) and mean squared error (MSE).

5 Analysis

The methods we developed and evaluated in the previous sections mainly serve as a means to carry out the analysis presented in this section. In particular, we applied the methods to all essays from the ICLE corpus (Granger et al., 2009). Based on their output, we analyze culture-specific argument structures in terms of what ADU flows learners use depending on three cultural aspects: the learners' *first language*, their *arguing proficiency*, and their *cultural context*. For each aspect, we also discuss the macro structures used in different cultures.

5.1 Differences across First Languages

One way to model culture is via the first language, that is, to assume all people with the same first language form one cultural group. While the ICLE corpus covers essays written by learners of 16 different first languages, we restrict our view to the five most representative ones: Chinese,[4] Tswana, Swedish, German and Italian.

ADU Flows Table 5 shows the five most frequent ADU flows in essays of learners of each considered first language. The essays from the European cultures (last three columns) comprise almost the exact same top ADU flows, with *premise (p)*, *claim (c)*, and *premise-premise (p-p)* as the top-3. In contrast, Chinese speakers largely start a paragraph with claims (*c*, *c-c*, and *c-p*), indicating a clear difference in argument structures compared European learners. Tswana speakers, finally, generally use more *premises* according to the output of our sentence-level ADU identifier.

Given that ADU flows are determined based on the ADUs within one paragraph each, the learners' paragraph splitting strategies may have affected the observed results. Table 6 shows statistics of paragraphs and their length across the cultures defined by the five languages. We see that the essays of all

[4] In this paper, we refer to both Chinese-Mandarin and Chinese-Cantonese as Chinese for simplicity.

Chinese		Tswana		Swedish		German		Italian	
c	4.2%	p-p	12.7%	p	3.8%	p	7.5%	p	11.3%
c-c	3.1%	p	11.5%	c	3.6%	c	7.1%	c	8.9%
c-p	2.5%	p-p-p	6.2%	p-p	3.3%	p-p	3.9%	p-p	4.3%
p	2.1%	c	4.2%	c-c	2.6%	n	2.9%	c-c	3.6%
n	2.0%	c-p	3.2%	p-p-p	2.3%	c-c	2.9%	n	3.5%

Table 5: First languages: The top-5 most frequent ADU flows and their occurrence in essays of learners from each of the five first languages. The letters *c*, *p*, and *n* stand for claim, premise, and none respectively.

	Chinese	Tswana	Swedish	German	Italian
# Essays	814	519	472	445	398
Paragraphs/essay	6.39	5.98	6.78	6.10	6.94
Sentences/parag.	4.46	3.25	4.52	4.39	3.33
Climactic	14%	7%	7%	6%	12%
Anti-Climactic	86%	93%	93%	94%	88%
Horizontal	58%	78%	68%	69%	71%
Vertical	42%	22%	32%	31%	29%

Table 6: First languages: The number of essays, average numbers of paragraphs per essay, and average number of sentences per paragraph in the essays of learners from the considered five languages. The lower part shows the proportion of essays that are climactic vs. anti-climactic as well as horizontal vs. vertical.

Argument Strength				Organization			
Low		High		Low		High	
p	7.5%	c	5.6%	p	7.9%	p	7.3%
c	5.2%	p	5.4%	c	5.9%	c	6.6%
p-p	4.4%	p-p	3.5%	p-p	3.4%	p-p	5.9%
n	3.3%	c-c	2.6%	n	3.2%	c-c	3.8%
p-p-p	2.2%	c-p	2.2%	c-p	1.6%	c-p	3.4%

Table 7: Arguing proficiency: The top-5 most frequent ADU flows and their occurrence in essay of learners with low and high arguing proficiency, according to our argument strength and organization scoring methods.

cultures have a similar number of paragraphs, likely due to the instructions on essay writing taught beforehand. Among the learners, Italians write the most with an average of 6.94 paragraphs, whereas Tswana speakers write the least: 5.98 paragraphs. Regarding the number of sentences in one paragraph, Italian and Tswana speakers write much fewer sentences compared to the other three languages in the table.

Macro Structures Additionally, we check for cultural differences in the macro-structure of the essays. On the hand, we counted how often they are *climactic* and how often *anti-climactic* (Suzuki, 2010). Climactic macro-structure refers to essays that have a writing style where the conclusion comes at the end (Suzuki, 2010). Statistically, English speakers generally tend to use an anti-climactic macro-structure where the conclusion appears at the beginning of articles. Computationally, we can see essays as climactic, if the extracted major claims are in the second half of the essay, and as anti-climactic otherwise.

On the other hand, we counted the numbers of *horizontal* and *vertical* essays (Suzuki, 2011). Horizontal macro-structure means the written arguments are not reason-based. In contrast, an essay is vertical, if the claims are supported by the premises (Suzuki, 2011). To distinguish the two cases, we assume that a claim is supported, if there is at least one premise appearing within the same paragraph. For example, the claim in Figure 2 is supported. With this in mind, we see an essay as having a horizontal macro-structure, if there are more claims being supported than the claims being unsupported.

With respect to the two kinds of macro-structures, Table 6 suggests that Tswana, Swedish, and German learners use fewer climactic essay constructions (6%–7%) than Chinese (14%) and Italian learners (12%). We also find that Tswana speakers use horizontal structures the most (78%), whereas Chinese speakers use them comparably little (58%).

Combining the results from Tables 5 and 6, we find a higher overall similarity between the argument structures of European cultures (Swedish, German, and Italian), matching intuition. Furthermore, among the three cultures, ADU flows and paragraph splitting strategies by Swedish and German speakers seem to be even closer. Our assumption is that the reason behind is these two languages belong to Germanic languages, whereas Italian has an entirely Roman origin.

5.2 Differences across Arguing Proficiencies

While we observed differences between learners of different first languages, they may partly also result from varying arguing proficiencies between the groups of learners. To further investigate this direction, we study ADU flows across proficiencies. In particular, we divided the essays based on their quality into two groups in two ways, once based on the argument strength scores and once based on the organization scores predicted by our methods. The

	Arg. Strength		Organization	
	Low	High	Low	High
# Essays	3 498	2 589	982	5 103
Paragraphs/essay	7.35	7.49	11.58	6.61
Sentences/paragraph	2.67	2.73	1.72	3.02
Climactic	10%	12%	11%	11%
Anti-Climactic	90%	88%	89%	89%
Horizontal	66%	57%	69%	61%
Vertical	34%	43%	31%	39%

Table 8: Arguing proficiency: The number of essays, average numbers of paragraphs per essay, and average number of sentences per paragraph in the essays of learners of different arguing proficiency. The lower part shows the proportion of essays that are climactic vs. anti-climactic as well as horizontal vs. vertical.

Argument Strength				Organization			
Low		High		Low		High	
I-B4-C	13.6%	I-B3-C	16.3%	I	13.4%	I-B3-C	17.1%
I-B3-C	13.2%	I-B4-C	14.4%	I-C	10.7%	I-B4-C	16.5%
I-B5-C	9.6%	I-B5-C	9.9%	I-B10-C	4.8%	I-B5-C	11.5%
I-B2-C	8.7%	I-B2-C	7.9%	I-B-C	4.1%	I-B2-C	9.8%
I-B6-C	6.4%	I-B6-C	6.5%	I-B11-C	3.3%	I-B6-C	7.6%

Table 9: Arguing proficiency: The top-5 most frequent paragraph flows and their occurrence in essays of low and high proficiency learners, according to our argument strength and organization scoring methods. I, B, and C mean *Introduction*, *Body*, and *Conclusion*, respectively. The number after of B means the number of paragraphs having the body labels.

essays that scored above or equal to the average scores (2.71 and 2.98, respectively) were classified as more proficient, the others as less proficient.

ADU Flows Table 7 shows the top-5 ADU flows written by learners of different arguing proficiency. In terms of *organization*, both groups share very similar patterns except for the fourth most frequent ADU flows (*n* vs. *c-c*). The flow *n* indicates that less proficient learners seem more prone to use non-argumentative text units. The results based on the *argument strength* scores reveal that the less proficient learners state premises more often than the more proficient ones (7.5% vs. 5.4%). Also for this quality dimension, we observe that less proficient learners resort more often to non-argumentative text units.

Macro Structures Table 8 presents statistics of the essays written by the two groups of learners. We find that, in terms of *argument strength*, the average number of paragraphs in an essay (7.35 and 7.49) and the average number of sentences in a paragraph (2.67 and 2.73) are very similar between writers of different proficiencies. However, in terms of *organization*, more organized essays tend to have notably fewer paragraphs (6.61 as opposed to 11.58), but much more sentences in one paragraph (3.02 as opposed to 1.72). This suggests that a good paragraph splitting strategy is key to better organization, while there is no clear clue how it affects argument strength.

Analyzing macro-structures, we also see that the proportions of climactic and anti-climactic essays are very similar for different proficiencies, both for argument strength and for organization. In terms of horizontal or vertical structures, more proficient learners seem to use more vertical structures (43% and 39%, respectively) in these two argument quality dimensions than less proficient ones (34% and 31%, respectively).

In Table 9, finally, we investigate the top-5 most frequent *paragraph flows*. A paragraph flow is here defined as a sequence of paragraph labels identified by the method, which we used for the corresponding feature in Section §3.2. We observe that less proficient writers in organization tend to write either too many (like 10 or 11) or very few (1 or even 0) body paragraphs. This again suggests that less proficient writers miss proper paragraph-splitting skills. For the argument strength, we find that both high and low proficiency writers have similar patterns. Note that, given that the paragraph labeling method may label the paragraphs incorrectly, we cannot say whether both high and low-proficiency learners split their essays in the same way into paragraphs. However, the results tell us that paragraph labels are not a clear feature to distinguish between essays having weaker and stronger argument strength.

5.3 Differences across Cultural Contexts

Another way to model culture is to split learners by whether they come from a high- or low-context culture. According to Hall (1976), “high context transactions feature pre-programmed information that is in the receiver and in the setting, with only minimal information in the transmitted message. Low context transactions are the reverse”. Zou (2019) sorts 15 languages from the lowest context culture to the highest. Since not all the languages in ICLE can be found in the sorted list, we select *Chinese* and *Japanese* to represent the high-context

High Context		Low Context	
claim	3.5%	premise	7.2%
claim-claim	2.6%	claim	6.7%
claim-premise	2.0%	premise-premise	4.0%
premise	2.0%	none	3.6%
none	1.9%	claim-claim	2.6%

Table 10: Cultural context: The top-5 most frequent ADU flows and their occurrence in essay of learners from high and low-context cultures.

Argument Strength				Organization			
High Context		Low Context		High Context		Low Context	
Low	High	Low	High	Low	High	Low	High
c	c	p	c	c	c	p	p
c-c	c-c	c	p	n	c-c	c	c
p	p-c	n	p-p	p	c-p	n	p-p
c-p	c-p	p-p	c-c	c-c	p	p-p	p-p-p
n	c-p-p	p-p-p	c-p	c-p	n	c-p	c-c

Table 11: Arguing proficiency and cultural context: The top-5 most frequent ADU flows in essays of learners from high- and low-context cultures, separately for essays of low and high proficiency, according to our argument strength and organization scoring methods.

cultures. For the low-context cultures, we select *German*, *Norwegian*, and *Czech*.

ADU Flows Table 10 shows the top-5 most frequent ADU flows in the high and low-context cultures. We find that learners from high-context cultures use more claims while low-context cultures use more premises in general. The reason behind this phenomenon may be that the pre-programmed information (premises in our case) is assumed to be known by the readers in the high-context culture. As a result, learners may, consciously or unconsciously, omit premises in their arguments.

Table 11 presents combined results for language proficiency and contextual cultures. In terms of the former, non-argumentative text units more frequently appear in the essays by less proficient learners from both cultural groups. The top ADU flow of high-context cultures is just a single claim (*c*), irrespective of the proficiency level. In contrast, both learners from low-context cultures tend to use more premises irrespective of proficiency.

Macro Structures Table 12 shows the macro-structure usage in the high and low-context cultures. We note that there is a tendency for high-context cultures to use more climactic (12% vs. 8%) and vertical (45% vs. 39%) structures in their writings. These findings fit the findings of Suzuki (2010) and Suzuki (2011). However, we point out that the majority of the macro-structure in our dataset is still anti-climactic and horizontal. The difference between the high and low context does not change this majority.

	High Context	Low Context
# Essays	1 348	1 002
Paragraphs/essay	6.02	7.02
Sentences/paragraph	5.26	5.06
Climactic	12%	8%
Anti-Climactic	88%	92%
Horizontal	55%	61%
Vertical	45%	39%

Table 12: Cultural context: The number of essays, average numbers of paragraphs per essay, and average number of sentences per paragraph in the essays of learners from high and low-context cultures. The lower part shows the proportion of essays that are climactic vs. anti-climactic as well as horizontal vs. vertical.

	Argument Strength				Organization			
	High ctxt.		Low ctxt.		High ctxt.		Low ctxt.	
	Low	High	Low	High	Low	High	Low	High
Climactic	11%	16%	7%	9%	5%	13%	6%	8%
Anti-Clim.	89%	84%	93%	91%	95%	87%	94%	92%
Horizontal	56%	49%	66%	54%	44%	55%	54%	64%
Vertical	44%	51%	34%	46%	56%	45%	56%	36%

Table 13: Arguing proficiency and cultural context: The proportion of essays that are climactic vs. anti-climactic as well as horizontal vs. vertical from high- and low-context cultures, separately for essays of low and high proficiency, according to our argument strength and organization scoring methods.

Finally, Table 13 analyzes the macro-structures considering both the language proficiencies and contextual cultures. It can be seen that most essays use an anti-climactic structure. For high-context cultures, learners of high proficiency use notably more climactic structures than those of low proficiency, both for argument strength (16% vs. 11%) and for organization (13% vs. 5%). For low-context cultures, there is a similar tendency, but with smaller differences (9% vs. 7% and 8% vs. 6%, respectively).

In terms of horizontal and vertical structures, we observe fewer horizontal ones in essays with higher argument strength than in those with lower argument strength for both cultural groups. The low-proficiency learners in low-context cultures use the most horizontal structures (66%) within the argument strength table block. In contrast, we observe an opposite situation in organization: writers use more horizontal structures in higher organization essays than in lower ones for both cultural groups.

6 Conclusion

This study aims to advance the understanding of language learners' argumentation with respect to cultural differences. To investigate argument structures in learner essays, we have built models for ADU identification and quality scoring, aiming at analyzing all ICLE essays. The results reveal differences and similarities of argument structures across English learners from different cultural backgrounds and proficiency levels.

The empirical findings from this study make two significant contributions to educational applications. First, argumentation technology can be of effective assistance in reducing the manual annotation workload as well as in expanding the research scope. Second, the analysis helps gain a comprehensive understanding of the argument structures produced by learners from different language backgrounds. It appears that culture would have a substantial influence on learners' argumentation patterns in terms of argument strength and organization. Our preliminary findings could be a doorway to the intercultural understanding of language learners' argument structures. For example, future research could usefully explore appropriate instructional approaches to help learners from different cultural backgrounds.

7 Limitations

While we provide many interesting findings in this paper, we are aware that there are several limitations in our study.

First, our analysis is based on the results of our ADU identification and quality scoring methods. More advanced models would be able to extract possible underlying patterns. It is likely that the top-5 ADU flows of each culture could be different from those retrieved in the current study.

Moreover, we notice that other factors other than mother tongue languages could play a vital role in the analysis of learners' argumentation structures. For example, the *first foreign language* or the *second language used at home*, both available in the ICLE dataset, could also influence the cultural backgrounds of the learners. These language usages may let them argue differently. However, in this study we only limit our view to their native language. Future studies can utilize more meta information of learners in order to figure out more cultural differences from other perspectives.

Last but not least, we do not distinguish languages spoken by multiple countries, e.g., German spoken in Germany and Switzerland. There could be some subtle differences in their argumentation strategies as well. In this paper, we assume that the language used in different countries share similar patterns regardless of where they are from. In the future, researchers can do further analyses by zooming in on these differences.

8 Ethical Statement

Our study can raise a few potential ethical concerns, as discussed in the following.

First of all, we show statistics of argument microstructures and macro-structures of different language groups. The results are not meant to be used to interpret that some cultural groups are better than others in any sense. Instead, the differences are a signal for understanding different cultural groups. While communicating with other people (e.g., in writing assistance), knowing the characteristics of their culture helps better understand them or what they may struggle with in expressing arguments. For example, knowing that low-context cultures expect many more premises in a statement, a speaker from a high-context culture can adjust the arguing strategies accordingly.

Secondly, our results should not be used to interpret that English learners from some cultural groups are good at arguing while some do not. We can conclude that some cultures use similar strategies to other cultures, and some cultures have their own strategies. While teaching languages, the results give hints for instructors on how to teach students accordingly. While designing argument mining models, the cultural group of the writers could be used as a feature in the models as well. Such applications of argument mining are expected to build on our findings.

Finally, we should be aware that the findings are based on whole cultural groups but not on individuals. We should not over-generalize or even stereotype people from different cultures in any situation. Still, people from a low-context culture may argue in the way that they are from a high-context culture. Any future research and application in this context should be aware of the individual differences.

Acknowledgments

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center “On-The-Fly Computing” (SFB 901/3) under the project number 160364472.

References

Esmaeel Abdollahzadeh, Mohammad Amini Farsani, and Maryam Beikmohammadi. 2017. Argumentative writing behavior of graduate efl learners. *Argumentation*, 31(4):641–661.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Fei-Wen Cheng and Yueh-Miao Chen. 2009. Taiwanese argumentation skills: Contrastive rhetoric perspective. *Taiwan International ESP Journal*, 1(1):23–50.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward T Hall. 1976. Beyond culture. garden city. *NY: Anchor*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Robert B Kaplan. 1966. Cultural thought patterns in inter-cultural education. *Language learning*, 16(1-2):1–20.

Il-Hee Kim, Richard C Anderson, Brian Miller, Jongseong Jeong, and Terri Swim. 2011. Influence of cultural norms and collaborative discussions on children’s reflective essays. *Discourse Processes*, 48(7):501–528.

Kyeongja Kim. 1997. A comparison of rhetorical styles in korean and american student writing. *Intercultural Communication Studies*, 6:115–150.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Donghong Liu and Fang Wan. 2020. What makes proficient writers’ essays more persuasive? A Toulmin perspective. *International Journal of TESOL Studies*, 2(1):1–13.

Xinghua Liu and Clare Furneaux. 2014. A multidimensional comparison of discourse organization in english and chinese university students’ argumentative writing. *International Journal of Applied Linguistics*, 24(1):74–96.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jin Kyung Paek and Yusun Kang. 2017. Investigation of content features that determine korean EFL learners’ argumentative writing qualities. *English teaching*, 72(2):101–122.

Daria Panina and Maya Kroumova. 2015. Cross-cultural communication patterns in computer mediated communication. *Journal of International Education Research*, 11(1):1–6.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages

543–552, Barcelona, Spain (Online). Association for Computational Linguistics.

Anushka Prakash and Harish Tayyar Madabushi. 2020. Incorporating count-based features into pre-trained models for improved stance detection. In *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 22–32, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Annotating argumentative structure in English-as-a-foreign-language learner essays. *Natural Language Engineering*, pages 1–27.

Jingjing Qin and Erkan Karabacak. 2010. The analysis of toulmin elements in chinese efl university argumentative writing. *System*, 38(3):444–456.

T Senthamarai and MR Chandran. 2015. Context in communication: A linguistic study of the interaction between the chinese and the indians in chennai, india. *Journal of Research in Humanities and Social Science*, 3(12):32–35.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Shinobu Suzuki. 2010. Forms of written arguments: A comparison between japan and the united states. *International Journal of Intercultural Relations*, 34(6):651–660.

Shinobu Suzuki. 2011. Trait and state approaches to explaining argument structures. *Communication Quarterly*, 59(1):123–143.

Sowmya Vajjala. 2018. Automated assessment of nonnative learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 1680–1691. The COLING 2016 Organizing Committee.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4063–4077, Online. Association for Computational Linguistics.

Yumei Zou. 2019. A study on english writing pattern under the impact of high-context and low-context cultures. In *5th International Conference on Arts, Design and Contemporary Education (ICADCE 2019)*, pages 758–762. Atlantis Press.