# 英語初学者エッセイの議論マイニングコーパスの作成と分析

川原田 将之 † 平尾 努 § 内田 渉 † 永田 昌明 §
† 株式会社 NTT ドコモ §NTT コミュニケーション科学基礎研究所
{masayuki.kawarada.vw, uchidaw}@nttdocomo.com
{tsutomu.hirao.kp, masaaki.nagata.et}@hco.ntt.co.jp

## 概要

議論マイニングとは，筆者の主張や根拠を特定し，それらの論理構造を明らかにするタスクである．学生の書いたエッセイを対象とした議論マイニングのコーパスとしては，Persuasive Essay Corpus (PEC) [1] が存在する．しかし，PEC は，留学生や英語を母国語とした高校生が書いたエッセイを基に作成されており，英語を外国語として学び始めた学習者 (初学者) が書く英文とは，記述されている英語の習熟度に差が存在する．本稿では，初学者が書いたエッセイに対して，PEC と同様のアノテーションを行い，新たに議論マイニングコーパスを作成した．複数名で作成したコーパスに対して一致率を算出し，PEC との比較分析を行った結果を報告する．

## 1 はじめに

英語を外国語として学ぶ者にとって，論理構造が明確な英文が書けることは，重要なスキルの一つである．学習者が書いた英文に対して，自動で論理構造の明確さに対する採点や修正箇所のフィードバックを行うようなシステムがあれば，英語学習に役立てることができ，有用である．このようなシステムを実現するためには，学習者が書いた英文を解析し，論理構造を明らかにする必要がある．

議論マイニングとは，筆者の主張や根拠を特定し，それらの論理構造を明らかにするタスクである．教育分野における議論マイニングのコーパスとしては，Persuasive Essay Corpus (PEC) [1] が存在する．PEC の英文は，留学生や英語を母国語とした高校生によって書かれたものであり，英語を外国語として学習する中学生や高校生が書くような英文は含まれていない．PEC におけるアノテーションのガイドラインも，意味上の段落であるパラグラフに分けて書かれているエッセイを前提としている．

一方で，初学者が書くエッセイは，パラグラフを意識して書かれていないといった特徴が存在する．このような特徴を持つ文書を対象にした議論マイニングの研究は少なく [2]，英語の習熟度が高い筆者が書いたエッセイと比較する研究は，今まで行われてこなかった．

そこで本稿では，PEC 作成におけるアノテーションガイドラインを用いて，初学者が書いたエッセイの議論マイニングコーパスを作成し，PEC との比較分析を行った結果を報告する．複数名で作成した議論マイニングコーパスに対して，アノテーション一致率を算出したところ，PEC と比較して，一致率は著しく低下した．一致率を上げるためのアノテーション方法の改善に向けて詳細な分析を行った．ごく少数のエッセイにおいては，筆者が意識してパラグラフを設けて作成されており，そのようなエッセイに限れば，アノテーション一致率は比較的高く，初学者によるエッセイに対するアノテーション方法の変更の方向性が示唆された．

## 2 関連研究

議論マイニングに関する研究は，ニュース記事を対象としたもの [3] や生物学の論文を対象としたもの [4, 5] など，様々な分野で行われている．教育分野における議論マイニングのコーパスを作成した研究としては，Stab ら [6] の研究がある．彼らは，エッセイ投稿用の web サイトから収集した 90 エッセイに対して，Component と呼ばれる筆者の主張が含まれる部分の抽出を行った後，Component 同士の関係を表す Relation の抽出を行っている．これに引き続き，Stab [1] らは，Stab ら [6] のアノテーション方法に修正を加えた方法で，PEC を作成した．PEC には，80 のテストデータと 322 の学習データが含まれており，テストデータに対してアノテーション一致率を計算している．

本研究では，Stab [1] らの研究を基に初学者のエッ

セイに対する議論マイニングコーパスを作成する．

# 3　議論マイニングコーパスの作成

先行研究 [1] との比較を行うために，初学者の書いたエッセイに対して，同じ方法でアノテーションされた，議論マイニングコーパスの作成行う．

## 3.1　エッセイの作成

エッセイの作成を行うために，筆者によって意見が分かれるような 250 のトピックを用意した．初学者[1]に，これらのトピックから．100 語から 150 語程度のエッセイを作成してもらった．各トピックに対して，10 エッセイを作成し[2]，合計で 2,500 エッセイを作成した．指定語数が 100 語から 150 語と短く，依頼を行う際に，パラグラフに分けて書くように指示はしていないため，パラグラフ区切りが含まれているエッセイは 102 例のみであった．

作成した 2,500 エッセイから，ランダムに 100 エッセイを抽出し，これらのエッセイに対してアノテーション作業を行った．抽出した 100 エッセイの中でパラグラフが存在しているものは 5 例だった．

抽出したエッセイの統計情報と PEC の統計情報を比較したものを付録 A の表 3 に示す．

## 3.2　アノテーション方法

Stab ら [1] のアノテーション方法では，3 段階に分けてアノテーションが行われる．まず，アノテータにトピックとエッセイを渡し，エッセイに書かれている内容を把握してもらう．次に，筆者の主張が含まれる要素である Component の抽出を行う．Component は，MajorClaim, Claim, Premise の 3 種類が存在し，Component の抽出は，文よりも小さい単位でで行われるため，抽出する境界の決定も必要である．Component の抽出は，MajorClaim, Claim, Premise の順に行う．最後に，抽出した Component 同士の関係を表す Relation の抽出を行う．Relation は，互いに同じ立場の関係にある Support と対立した立場の関係にある Attack が存在する．

Stab ら [1] の方法では，エッセイにパラグラフが存在することを前提に，各パラグラフ内で相対的に Component と Relation の抽出を行っている．一方で，初学者のエッセイでは，パラグラフが存在しない場合が多く，エッセイ全体を見ながら抽出を行う必要がある．

## 3.3　アノテーション作業

アノテーション作業は．英語を母国語とする 3 名のアノテータで行った．まず，3 名には，先行研究 [1] のアノテーションガイドライン[3]を用いて，アノテーション方法の習得をしてもらった．その後，各アノテータは，それぞれ 100 エッセイに対してアノテーションを行った．

今回作成したエッセイは，PEC と比べて，エッセイ作成者の英語の習熟度が高くないため，アノテータが理解できない場合や，そもそもトピックに対応した主張が含まれていない場合が想定される．これらに対応するため，アノテータには，どこまでの作業ができるのかを判断してもらい，作業を完了できなかった場合は，その理由を記述してもらった．

表 1 に作成したエッセイに対して行ったアノテーションの例を示す．また，作成した議論マイニングコーパスについて，各アノテータが抽出した Component の数と付与した Relation の数，作業不可と判断されたエッセイの数を付録 A の表 4 に示す．

# 4　アノテーション一致率の算出方法

アノテータが可能だと判断した作業についての一致率を算出した後，Stab ら [1] と同様の方法で,Component の一致率と Relation の一致率をそれぞれ算出する．

## 4.1　作業判断についての一致率

本研究では，先行研究 [1] とは異なり，アノテータが作業不可と判断することも可能である．各エッセイに対して，3 名のアノテータが作業可能と判断したか ($t = 1$)，または，作業不可と判断したか ($t = 0$) を Fleiss [7] の $\kappa$ を用いて，アノテーション一致率を計算する．

## 4.2　Component の一致率

まず，文単位でアノテーション一致率を計算する．この方法では，アノテータ間で Component を抽出する境界が異なっていたとしても，抽出した Component が同一の種類であれば，一致していると評価される．文単位のアノテーション一致率の計

1）　本研究で初学者とは，TOEIC ® Listening & Reading Test で 500 点以下の点数かつ，英語を母国語としない者とした．
2）　ただし，1 人の初学者が同一のトピックについて複数のエッセイを書くことがないように調整した．
3）　https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422

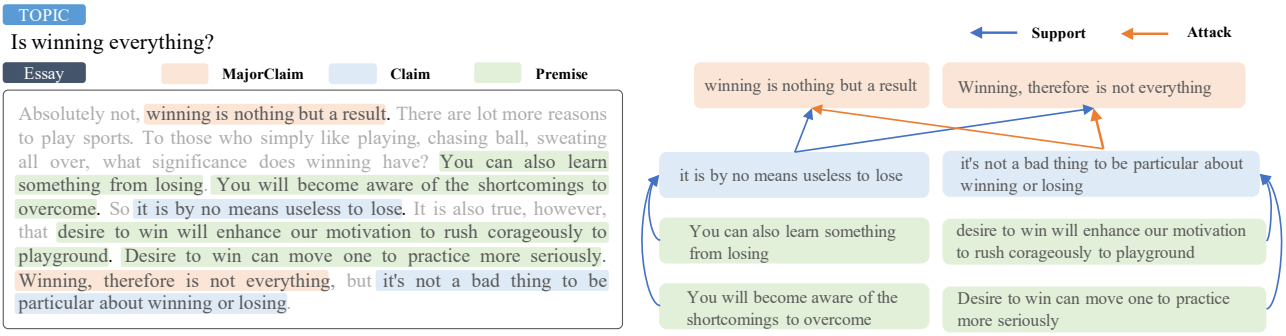


**図 1**　作成したエッセイに対して行った，Component（左）と Relation(右) のアノテーションの例

| | 先行研究 [1] | | | 本研究 | | |
|---|---|---|---|---|---|---|
| | OA | $\kappa$ | $\alpha_U$ | OA | $\kappa$ | $\alpha_U$ |
| MC | 0.979 | 0.877 | 0.810 | 0.872 | 0.530 | 0.497 |
| C | 0.889 | 0.635 | 0.524 | 0.685 | 0.247 | 0.232 |
| P | 0.916 | 0.833 | 0.824 | 0.683 | 0.292 | 0.271 |

| | 先行研究 [1] | | 本研究 | |
|---|---|---|---|---|
| | OA | $\kappa$ | OA | $\kappa$ |
| S | 0.923 | 0.708 | 0.906 | 0.350 |
| A | 0.996 | 0.737 | 0.984 | 0.072 |

**表 1**　Component(上) と Relation(下) の抽出に関する先行研究 [1] と本研究のアノテーション一致率の比較．OA は，Observed Agreement を表す．MC，C，P，S，A は，それぞれ，MajorClaim，Claim，Premise，Support，Attack を表す．

| | MC | C | P | 合計 |
|---|---|---|---|---|
| アノテータ A | 1 | 0 | 1 | 2 |
| アノテータ B | 8 | 0 | 5 | 13 |
| アノテータ C | 1 | 1 | 4 | 6 |

**表 2**　3 名のアノテータが各作業段階において，作業不可と判断したエッセイの数．MC，C，P は，それぞれ，MajorClaim，Claim，Premise を表す．

算には，観察された一致度 (Observed Agreement) と Fleiss [7] の $\kappa$ を用いた．観察された一致度は，単純に全体の中で 3 名の作業者の一致した割合を測る指標であり，$\kappa$ は，観察された一致度から偶然の一致を除いた時の一致率を測る指標である．

次に，Component の境界も考慮した評価を行うために，Krippendorff [8, 9] の $\alpha_U$ の算出を行う．この算出方法では，各エッセイを単語に分割し，[4] エッセイ全体を単語列と見なした上で，アノテータが抽出した Component の境界とその種類の一致率を計算する．そして，算出した各エッセイの一致率の平均値を取ることでアノテーション一致率とする．

### 4.3　Relation の一致率

先行研究 [1] に従い，Component を含んだ 2 文間の Relation の一致率を算出する．あるパラグラフ内に含まれる文の総数を $n$ とすると，パラブラフ内の文はそれぞれ，$s_1, \ldots, s_n$ と表され，その中の任意のペアは，$p = (s_i, s_j)$ と表される．ここで，$0 \leq i, j \leq n$ かつ $i \neq j$ である．全エッセイに含まれるペアの総数 $N$ うち，アノテータ間で Relation の種類が一致している割合からアノテーション一致率を計算する．ただし，パラグラフの区切りが存在しないエッセイでは，$n$ はそのエッセイに含まれる文の数と一致する．アノテーション一致率には，Observed Agreement と Fleiss [7] の $\kappa$ を用いた．

## 5　分析結果

3 名のアノテータによるアノテーション一致率を算出するため，1 人でも作業が不可と判断したエッセイは除外した．アノテーションが完了した 81 エッセイに対して，前節の方法でアノテーション一致率を算出し，先行研究 [1] との比較結果を示す．

### 5.1　作業不可と判断されたエッセイに対する分析

100 エッセイのうち，除外されたエッセイは，19 エッセイであり，3 人のアノテータ全員が作業不可と判断したエッセイは存在しなかった．アノテータ 3 名の作業判断についての一致率を計算したところ，$\kappa = 0.087$ が得られた．また，除外した 19 エッセイのうち，13 エッセイは，アノテータ B のみが作業不可と判断したエッセイであったため，アノテータ A とアノテータ C の一致率を算出したところ，$\kappa = 0.2647$ であった．これらの結果を見ると，作業判断についての一致率は低く，作業判断はアノテータによって異なることがわかる．

次に，除外されたエッセイがどの作業段階で作業不可と判断されたのかについて考察する．アノテーション作業において，Relation を抽出する段階で作業不可と判断された例は存在しなかった．これは，Component の抽出の際には，他の Component との対

4)　単語分割には，NLTK 3.7 を用いた．

応を考えながら行う必要があるため，Component の抽出ができれば，Relation の抽出も行えるからだと考えられる．3 名のアノテータが各作業段階において，作業不可と判断したエッセイの数を表 2 に示す．MajorClaim の抽出を行う段階で，作業不可判断されたエッセイは，10 例であった．作業不可と判断した理由としては，「トピックに対応した主張が含まれていない」や「エッセイ自体が不明瞭で主張が理解できない」が多かった．また，Premise の抽出において，10 例が作業不可と判断されていた．その理由としては，「Premise に該当する箇所がエッセイ中に存在しない」がほとんどであった．これは，初学者のエッセイは，主張を書くことができたとしても，それらの因果関係が読者に伝わりづらい英文であるためだと考えられる．

### 5.2 アノテーション一致率に関する分析

Component と Relation のそれぞれにおいて，先行研究と本研究のアノテーション一致率を比較したものを表 1 に示す．まず，Component について見てみると，$\kappa$，$\alpha_U$ の値は，MajorClaim で $\kappa = 0.530$，$\alpha = 0.497$ であった．MajorClaim の一致率が比較的高くなった理由としては，一般的に MajorClaim は文書の最初と最後に出てくる場合が多く，エッセイ作成者の意図が分かりづらくても，文の位置からある程度は予測できるからだと考えられる．先行研究では，Premise の一致率は $\kappa$，$\alpha_U$ 共に高いが，本研究では低く，$\kappa = 0.247$ と $\alpha_U = 0.497$ だった．これは，初学者のエッセイには，接続詞の誤りが多く含まれ，主張と理由の判別が難しいためだと考えられる．

次に，Relation について比較を行う．Support についての $\kappa$ の値は，$\kappa = 0.350$ であり，先行研究と比べると高くないものの，ある程度の一致率が得られた．一方で，Attack についての一致率は低く，$\kappa = 0.072$ であった．この理由として，初学者のエッセイは主張の立場がわかりづらく，アノテータが判別するのが難しかったためだと考えられる．

### 5.3 パラグラフの影響に関する分析

パラグラフの有無がアノテーション一致率に与える影響について調べるために，これらを 2 つの群に分けた後，Component の一致率を比較する．比較には，$\alpha_U$ を用いた．パラグラフが存在するエッセイと存在しないエッセイに対して $\alpha_U$ を比較したものを図 2 に示す．　MajorClaim，Claim，Premise の全

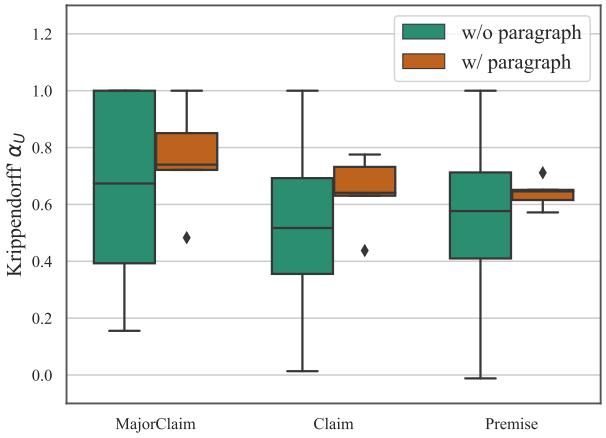


**図 2**　パラグラフが存在するエッセイ (w/ paragraph) と存在しないエッセイ (w/o paragraph) に対する $\alpha_U$ の比較

てにおいて，パラグラフがある場合の方がエッセイのアノテーション一致率が高いことがわかる．これは，パラグラフが存在することにより，エッセイ作成者が意図した，意味上の区切りをアノテータが理解しやすくなるためであると考えられる．特に，本研究のような短いエッセイの場合，MajorClaim の含まれるパラグラフは 1 文であることが多く，パラグラフがないものに比べて作業が容易であったと考えられる．ただし，パラグラフが存在するエッセイと存在しないエッセイの間には，サンプル数に大きな偏りがあるため，同程度のサンプル数における比較は，今後の課題である．

これらの結果から，初学者のアノテーション作業を行う際には，パラグラフに分割可能かを判断する作業を追加するなど，作業工程の検討を行う必要があると考えられる．

## 6 おわりに

本研究では，初学者が書いたエッセイに対して，議論マイニングのアノテーションを行った．複数のアノテータで作成した議論マイニングコーパスについて，アノテーション一致率を算出したところ，先行研究 [1] よりも一致率が低い結果が得られた．そこで，パラグラフの区切りが存在するエッセイと存在しないエッセイについて，アノテーション一致率の比較を行ったところ，パラグラフが存在するエッセイの方が一致率が高い結果が得られた．今後は，アノテーション作業におけるパラグラフの影響について，より詳細な分析を行った後，初学者のエッセイに対するアノテーション方法の検討を行いたい．

# 参考文献


[1] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. **Computational Linguistics**, Vol. 43, No. 3, pp. 619–659, September 2017.

[2] Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. Parsing argumentative structure in English-as-foreign-language essays. In **Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 97–109, Online, April 2021. Association for Computational Linguistics.

[3] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 3433–3443, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[4] Nancy Green. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In **Proceedings of the First Workshop on Argumentation Mining**, pp. 11–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[5] Nancy Green. Identifying argumentation schemes in genetics research articles. In **Proceedings of the 2nd Workshop on Argumentation Mining**, pp. 12–21, Denver, CO, June 2015. Association for Computational Linguistics.

[6] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**, pp. 1501–1510, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[7] J.L. Fleiss, et al. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Vol. 76, No. 5, pp. 378–382, 1971.

[8] Klaus Krippendorff. **Content Analysis: An Introduction to Its Methodology (second edition)**. Sage Publications, 2004.

[9] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

# A　データ作成時の統計情報

| | 先行研究 [1] | 本研究 |
|---|---|---|
| | 全エッセイの合計 | |
| 文数 | 7,116 | 743 |
| 単語 | 147,271 | 13,704 |
| パラグラフ数 | 1,833 | 14 |
| | 各エッセイあたりの平均 | |
| 文数 | 18 | 7.43 |
| 単語数 | 366 | 137.04 |
| パラグラフ数 | 5 | 0.14 (2.8) |

**表 3**　先行研究 [1] と本研究でアノテーションを行うエッセイの統計情報の比較．括弧内の値は，パラグラフが存在した 5 例のエッセイのみの平均値を表す．

| | Component の数 | Relation の数 | 作業不可の数 |
|---|---|---|---|
| アノテータ A | 705 | 551 | 2 |
| アノテータ B | 510 | 410 | 14 |
| アノテータ C | 564 | 459 | 5 |

**表 4**　各アノテータか抽出した Component の数と Relation の数，および，作業不可と判断したエッセイの数 (作業不可の数)．