ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 318-347

Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach

Sathena CHAN* Manoranjan SATHYAMURTHY** Chihiro INOUE***
Michael BAX**** Johnathan JONES***** John OYEKAN******

Abstract

In recent years, large-scale language test providers have developed or adapted automated essay scoring systems (AESS) to score L2 writing essays. While the benefits of using AESS are clear, they are not without limitations, such as over-reliance on frequency counts of vocabulary and grammar variables. Discourse competence is one important aspect of L2 writing yet to be fully explored in AEE application. Evidence of discourse competence can be seen in the use of Metadiscourse Markers (MDM) to produce reader-friendly texts. The article presents a multidisciplinary study to explore the feasibility of expanding the construct representation of automated scoring models to assess discourse competence in L2 writing. Combining machine learning, automated textual analysis and corpus-linguistic methods to examine 2000 scripts across two tasks and five proficiency levels, the study investigates (1) in addition to frequency and range, whether accuracy of MDM is worth pursuing as a predictive feature in L2 writing, and (2) how identification and classification of MDM use might be fed into developing an automated scoring model using machine learning techniques. The contributions of this study are three-fold. Firstly, it offers valuable insights within the context of Explainable AI. By integrating MDM usage and accuracy into the scoring framework, this research moves beyond frequency-based evaluation. This study also makes significant contributions to the current understanding of L2 writing development that even lower-proficiency learners exhibit evidence of discourse competence through their accurate use of MDMs as well as their choice of MDMs in response to genre. From the perspective of expanding the construct representation in automated scoring systems, this study provides a critical examination of the limitations of many AEE models, which have heavily relied on vocabulary and grammar features. By exploring the feasibility of incorporating MDMs as predictive features, this research demonstrates the potential for construct expansion of L2 AEE. The results would support test providers in developing competence tests in various contexts and domains including manufacturing, medicine and so on.

Keywords: L2 Writing, Metadiscourse Markers, Automated Essay Scoring, Large Language Models

* Assoc. Prof. Dr., University of Bedfordshire, CRELLA, UK, sathena.chan@beds.ac.uk, ORCID ID: 0000-0002-7852-6737
** Researcher, University of Oxford, UK, manosathya98@gmail.com, ORCID ID: 0000-0001-8928-2689
*** Assoc. Prof. Dr., University of Bedfordshire, CRELLA, UK, chihiro.inoue@beds.ac.uk, ORCID ID: 0000-0003-1927-6923
**** Researcher, Weblingua LTD, UK, michael@textinspector.com, ORCID ID: 0000-0002-2753-1990
***** Dr., University of Bedfordshire, CRELLA, UK, johnathan.jones@beds.ac.uk, ORCID ID: https://orcid.org/0000-0003-4158-7971
****** Senior Lecturer, University of York, UK, john.oyekan@york.ac.uk, ORCID ID: 0000-0001-6578-9928

To cite this article:
Chan, S., Sathyamurthy, M., Inoue, C., Bax, M., Jones, J., & Oyekan, J. (2024). Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 318-347. https://doi.org/10.21031/epod.1531269

Received: 13.08.2024
Accepted: 16.12.2024

Introduction

In recent years, large-scale language test providers have developed or adapted automated essay scoring systems (AESS) to score second language (L2) writing essays. For example, Educational Testing Service uses Natural Language Processing based e-rater® Scoring Engine and Pearson uses Intelligent Essay Assessor™ through a combination of Latent Semantic Analysis (LSA) and other methods. While the benefits of incorporating AEE applications in the scoring systems are clear, they are not without limitations. Early systems were criticized for their over-reliance on frequency counts of vocabulary and grammar variables (Chapelle and Chung, 2010). Current state-of-art AESS have incorporated scoring features such as content and organization. However, discourse competence as one important aspect of L2 writing is yet to be fully explored in AESS. Discourse competence "concerns the ability to design texts, including generic aspects like thematic development and coherence and cohesion as well as … cooperative principles and turn-taking" (CoE, 2018, p.138). In writing, evidence of discourse competence can be seen in the use of metadiscourse markers (MDM) to produce reader-friendly texts. Such competence is typically expected from higher-proficiency L2 writers learners, especially at the CEFR B2 level or onwards (CoE, 2018), when they have mastered linguistic accuracy and basic writing skills Nevertheless, in the increasingly multicultural contexts we live in, discourse competence which underpins effective communication is relevant to L2 learners across the proficiency spectrum, arguably more so for lower-proficiency learners who need to build meaningful connections and achieve educational/professional goals. The article presents a multidisciplinary study to explore the feasibility of expanding the construct representation of AESS to assessing discourse competence in L2 writing. This would improve the way tests are developed and assessed across various contexts, domains and sectors including manufacturing, construction, medicine and so on thereby supporting low-skilled to highly skilled labor in these areas.

Use of MDMs in L2 writing tests

Metadiscourse markers (MDM) are defined in this study as "those aspects of the text which explicitly refer to the organization of the discourse or the writer's stance towards either its content or the reader" (Hyland, 2005, p. 109). The use of MDMs has two major functions. Firstly, skilled writers use MDMs to signal the organization of a text and provide cohesion between ideas in a text, e.g., to indicate conjunctive and/or additive, adversarial, causal and temporal relationships in the text (Schiffrin et al., 2001, p.55). Secondly, MDMs are used to state the attitude of the writer (Burneikaite, 2008). Skilled writers use MDMs to provide an explicit organizational structure within a text and to guide the reader to their attitude on the topic. Appropriate use of MDMs makes a text more reader-friendly, especially for L2 readers (Camiciottoli, 2003). Despite the importance of discourse competence in the development of L2 writing proficiency, especially when learners progress to CEFR B2 or upwards (CoE, 2001), evaluation of the use of MDMs in L2 writing is typically reduced to a holistic judgment of the number and/or range of cohesive devices used under the criterion of "cohesion and coherence" in human scoring schemes (for example see the Aptis Guide, 2019). This approach might be limited to reveal the nuanced developmental features of the use of MDMs by L2 writers.

We now review the previous studies on the use of MDMs in L2 writing. Most of these studies focused on the use of MDMs by upper-intermediate L2 writers, comparing their academic essays with those of L1 writers (e.g., Adel, 2006; Crompton, 2012; Hyland, 2005; Lee & Deakin, 2016). Their findings are clearly inconclusive and contradictory at times. Some studies found that higher-proficiency writers use more MDMs overall than lower-proficiency writers (Sanford, 2012). Others reported higher use of certain MDMs (such as endophoric markers and evaluative markers) among higher-proficiency writers (Burneikaitė, 2008). In contrast, some reported that higher-proficiency writers use fewer logical connectives than lower-proficiency writers but used a wider range of MDMs (Carlsen, 2010).

Only a handful of studies investigated the use of MDMs by L2 learners in standardized writing tests. In Knoch et al.'s (2014) study on TOEFL writing test, lower-proficiency writers used more MDMs overall

than more proficient writers. Bax et al. (2019) conducted the first large-scale study to examine L2 test takers' use of MDMs. 900 writing scripts produced by L2 test takers at CEFR B2-C2 levels were examined. They found that higher-proficiency writers used fewer MDMs but a significantly wider range of MDMs than lower-proficiency writers. Barkaoui (2016) investigated only interactional MDMs among repeaters of IELTS and found no significant effects in test taker group on the overall use of interactional MDMs. However, test takers scoring IELTS 6.0 (indicating CEFR level B2 according to test providers' information) tended to use more hedges and boosters but fewer self-mentions than did test takers with lower initial writing scores. Owen et al. (2021) expanded on Bax et al.'s work to include test takers from CEFR A levels. The results showed that each of the 13 MDM categories used in their study discriminated across at least one CEFR boundary. The overall deployment of MDMs changed significantly in transitioning from A0 to A2 levels and from B1 to C levels. The range of MDMs also rose across CEFR thresholds, with significant differences obtained across A1-A2 and A2-B1 thresholds. As a result, they argued that the use of MDMs should be operationalized separately from vocabulary (grammatical competence) as part of discourse competence (Bachman and Palmer, 2010).

These studies clearly show differences in frequency and range of MDMs used by L2 writers, indicating that increasing (or sometimes decreasing) use of MDMs may signal test takers' ability to manage textual and interpersonal complexity in discourse. However, the findings are inconclusive in at least two aspects. First, the direction of the relationship between the use of MDMs (frequency and range) and L2 writing proficiency is inconclusive. Second, differences in frequency and range of MDMs seem observable between some levels but not others. As a result, simple frequency and range counts of MDMs might not be the most suitable way of distinguishing between L2 writing proficiency levels, especially for writing tests which target multiple proficiency levels.

Potentials and challenges of Automated Essay Scoring Systems

Automated essay scoring systems (AESS) have become increasingly prevalent in the assessment of L2 writing. A range of lexical and some syntactic measures have been shown to consistently discriminate across score boundaries in large-scale testing. Lexical complexity can be measured in terms of rarity, variability and disparity (Jarvis, 2013). For example, word frequency counts in relation to threshold levels of vocabulary use based on various wordlists, e.g. English Vocabulary Profile, Academic Word List and New General Service List (Brezina & Glabasova, 2013) are commonly used in L2 writing AESS. However, most AESS rely heavily on frequency and range of lexical use based on frequency wordlists.

The performance of pre-trained transformers on various NLP tasks is well documented, however this does not necessarily translate to good out-of-the-box performance on all downstream tasks presented to the model (Lin et al., 2022). Currently pre-trained transformers have been used to obtain word embeddings; after which a classifier has been trained to perform our binary classification task. We can build upon the knowledge that the pre-trained transformer has learnt by fine-tuning the model using our labelled dataset. In one such fine-tuning method, we can alter parameters in a given number of layers in the transformer architecture that we wish to fine-tune, leaving parameters in all other layers untouched (Lialin et al., 2023). This can, depending on the level of fine-tuning, potentially be a reasonably resource consuming task; it is, however, capable of boosting performance for particular tasks.

Taken together, research is needed to explore whether AESS can be extended to detect frequency and range of MDM as measures of discourse competence in L2 writing and whether MDM accuracy can serve as a predictive feature in L2 writing.

Methods

Through a multidisciplinary approach combining machine learning, automated text analysis and corpus-linguistic methods, we investigated whether MDM accuracy is a predictive feature in L2 writing and the

feasibility of building an automated scoring model to identify use of MDMs and to distinguish between accurate and inaccurate use of MDMs. Three research questions guided this study:

RQ1: How do learners use MDMs across proficiency levels?

RQ2.1: To what extent can a transformer-based AI model classify correctly whether or not a word or a phrase is a MDM?

RQ2.2: To what extent can a transformer-based AI model classify correctly the accurate and inaccurate use of MDMs?

The Research Questions were addressed in two phases. Phase 1 involved human coding to identify and examine frequency and accuracy of use of MDMs by test takers taking a large-scale proficiency writing test, and to explore that use across a range of CEFR levels. Phase 2 involved use of machine learning of the human-coded data to investigate which machine learning algorithms could be used to develop an automated model to replicate expert judgement on detection and accuracy of MDMs.

Tasks and data set

The dataset for the study consists of 2,003 sample scripts from the corpus of Aptis candidates' writing. Aptis is a standardized multi-level English Proficiency Test. The Writing component of the Aptis test consists of four parts. Part 4 (Formal and Informal Writing) was used in the current study. Aptis writing is evaluated by trained and certified human examiners. Although Aptis employs a single-rating approach, different raters are assigned to each task, ensuring that multiple observations are made of a single candidate's response. The inter-rater reliability on benchmark Writing responses is at 0.97 (O'Sullivan, Dunlea, Spiby, Westbrook, and Dunn, 2020). Since Aptis is taken by candidates in different international contexts, candidates are allowed to use any standardized version of English (e.g., American, Australian, British, Singaporean) in the writing test as long as it is consistent, especially in the formal writing task.

The scripts used in this study were from the two email tasks in Part 4 of the Aptis writing test. The two tasks were thematically-linked. Task 1 required the candidates to write an email (40-50 words) to a classmate friend about a class cancellation in a cooking school as the teacher is going on a holiday. Task 2 required the candidates to write an email (120-150 words) to the manager of the cooking school to complain about the cancellation. They had 20 minutes to finish each task. Each script was operationally tagged with a CEFR level based on the candidate's test scores received on the task (as part of the standard test procedure in Aptis), and the breakdown of the numbers of scripts at the five CEFR proficiency levels is shown in Table 1.

Table 1

Numbers of scripts used for analysis in this study

	A1	A2	B1	B2	C	Total
Task 1	175	210	190	187	234	996
Task 2	173	206	197	193	238	1007

MDM Categories

We used the categories of MDM shown in Table 2 (Hyland, 2005, modified by Bax et al., 2019) (see Appendix 1 for the full list).

Table 2

Categories of MDM

	Category analyzed		Function	Examples
Textual metadiscourse	Logical connectives		Express semantic relation between main clauses	In addition / but / thus / and
	Frame markers:	Sequencing	Explicitly refer to discourse acts or text stages	Finally / to repeat / here, we try to
		Label Stages		
		Announce goals		
		Topic shift		
	Code glosses		Help readers grasp meanings of ideational material	Namely / such as / e.g. / i.e.
	Endophoric markers		Refer to information in other parts of the text	Noted above / see figure X
	Evidentials		Refer to source of information from other texts	According to X, … / 1990 / X argues that…
Interpersonal metadiscourse	Attitude markers		Expressing opinion of propositional content	I agree that… / X claims that…
	Hedges		Withhold writer's full commitment to statements	Might / perhaps / possible
	Relational markers		Explicitly refer to or build relationship with reader	Frankly / note that / as you can see…
	Person markers		Explicit reference to author	I / we / mine / our
	Emphatics		Emphasize force or certainty in message	Definitely / in fact / it is certain that…

Procedures for RQ1 (Use of MDM at different proficiency levels)

A total of 996 Tasks 1 and 1007 Tasks 2 scripts were manually coded using the procedures described below. The manual coding results were then used to build a labelled training dataset as the first step for developing a transformer-based AI model to identify and assess accuracy of MDM in RQ2.

Automated tagging of MDM and data cleaning

Text Inspector, a web-based tool allowing users to analyze features of texts, was used to provide an initial tagging of MDM using categories of Hyland's (2005) list. Adopted from the procedures used in Owen et al. (2021), we cleaned the tagged dataset as follows:

- Full stops and exclamation marks were removed, since the units of analysis were not sentences;
- Special symbols were removed or replaced with correct ones; and
- Spelling errors were corrected to improve the accuracy of automated classification.

During the initial coding of the dataset (i.e., 30% with over 100,000 words in total for each task), we found that more than half of the inaccurate uses of MDMs were spelling errors[1] (e.g., 102 out of 200 inaccurate MDMs in Task 1). A decision was made to correct them for two reasons. First, the focus of the study is about the frequency, range and accuracy of MDM use by L2 writers. As argued previously, this is related to their discourse competence (Bachman and Palmer, 2010) to signal the organization and/or the author's stance in a text for its reader rather than their ability to spell the markers correctly, which is typically assessed in relation to "vocabulary" in L2 writing. Secondly, inclusion of misspelled words would increase variation for the algorithms to accurately classify use of MDMs.

RQ1 Coding Procedures

The tagged scripts were then reviewed and coded manually for the use of MDMs by two researchers, following these procedures:

1) Adding any words and expressions to the list of MDMs that are suitable for the genre of email writing. As Hyland's (2005) list was devised based on journal articles, it does not include the full range of MDMs that were found in emails in the current study. For example, among frame markers in Hyland's list, examples expressions include 'here, we try to…' for announcing the goal of the piece of a text. However, this expression is unlikely to be used in emails; instead, we frequently observed 'I am writing this email to…' at the earlier part in emails, which need to be added to the list for this study. There were also more varied attitude markers such as 'disappointing/ disappointed' and 'happy' in the scripts than would be in journal articles. The list of additional MDMs can be found in Appendix 2.

2) Indicating any words or expressions tagged according to Hyland's list that do not serve as MDMs in the current data set. Related to the above point, there are some words and expressions that qualify as MDMs in journal articles, but not in emails. For example, the word 'next' is tagged as a MDM according to Hyland (2005), which signals the sequencing of texts in journal articles (e.g. 'Next, we examine…'). However, in the scripts in this study, 'next' is often used to say 'next week', which does not serve as a MDM in the simulated email texts. These non-applicable tags were identified and removed during coding.

3) Code dichotomously for the identification and accurate or inaccurate use of correctly-tagged (by Text Inspector) and newly identified MDMs (see Figure 1). Specifically, the two coders make decisions on two questions:

Q1: Is this a MDM? (1: yes, 0: no)

Q2: If it is a MDM, is it correctly used ? (1: yes, 0: no)

The coded data was used to address RQ1 (i.e. the frequency, range and accuracy of MDM use across proficiency levels) as well as serving a labelled set for training algorithms for RQs 2.1 and 2.2.

Due to the exploratory nature of this first study to develop AESS models to assess the use of MDMs by L2 writers, we sought a dichotomous instead of polytomous coding scheme regarding the accuracy of MDM because the latter would require a more complex model for machine learning (more will be discussed regarding the procedures for RQ2). Because of the nature of the dichotomous coding scheme, the inaccurate MDM use needed to be undoubtedly inaccurate, see examples below. In this study, the 0 codes (for inaccurate use) therefore largely represented grammatical errors that surround MDMs use (see Examples 1-3)

Example 1: when you return of <u>you</u> holiday ….(A1 script, relational marker)

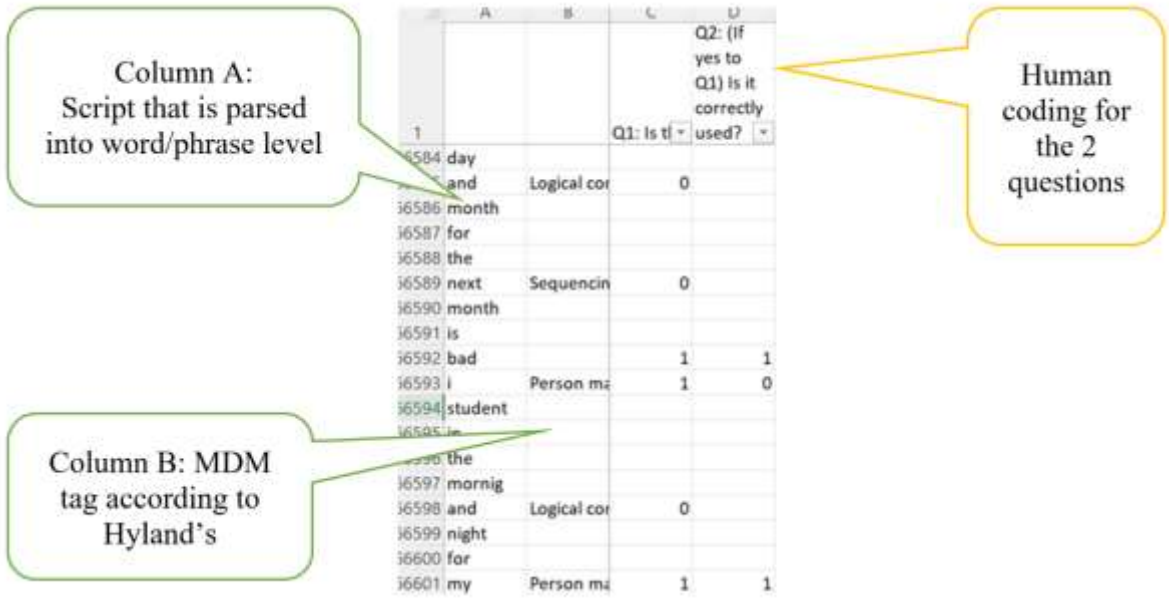
Example 2: 20th of the next moth fo <u>my</u> is the most ….(A2 script, relational marker)

[1] When a potential spelling error was identified and the spelt word exists in English, it was not corrected (for example, in the case of 'Thank your', "your" could have been misspelt (instead of "you"). But since "your" in itself is not misspelt, it was not corrected.

Example 3: Please let's know on the status … (A2 script, relational marker)

Figure 1

A screenshot of example human coding



As a result, the range of inaccurate MDM use is narrower in this study than what might usually be regarded as inaccurate use. Less appropriate use of MDMs (such as using a formal label stage when writing to a friend in informal email task (see Example 4) and using emphatics instead hedges when writing to a manager in the formal email task – see Example 5) was not coded as 0 (inaccurate).

Example 4: in conclusion, … (B1 script, label stage)

Example 5: I am feel really disappointed …(B1 script, emphatics)

To differentiate these developmental features of the discourse competence, a polytomous coding scheme, which was deemed inappropriate for this first study, would be required. The two coders double-coded 123 scripts per task, which makes up 10% of the data. After several rounds of discussions and re-coding, the (working) list of additional MDM for email writing (Appendix 2) was agreed and the exact agreement rate reached over 90% for both tasks (Task 1, Q1[MDM or not]: 96.4%, Task 1: Q2 [accurate use or not]: 96.1%; Task 2, Q1 (98.5%, Task 2, Q2: 98.3%). The coding reliability was deemed sufficiently high, and thus the two coders continued on to code two different sets of scripts (each batch containing 45% of the scripts) independently.

We report a descriptive summary of human coding in relation to the ratio of scripts where at least one MDM (irrespective of accuracy) in each MDM category across proficiency levels appeared for each task to show the general trend. Kruskal-Wallis tests were then run for the average ratio of accurately used MDMs across levels.

Procedures for RQ2 (Transformer-based model to classify use and accuracy of MDM)

To remind the reader, RQ2 aims to explore how a machine learnt automated scoring model could be applied to evaluate test taker's MDM use. This involved four stages: the experimental setup, production of word embeddings, automated classification using word embeddings, and finally improvement of the classifiers used in the classification task.

(1) Experimental Setup

The human coded scripts were used as a labelled dataset for this part of the study. For the purpose of this project, we considered each of the research questions, i.e., RQ2.1 and RQ2.2, as an individual binary classification task. The premise was that each word in our dataset can be labelled as a 1 or 0.

a. 1 – a word is a MDM or 0 – a word is not a MDM [RQ2.1]

b. 1 – it is accurately used or 0 – it is not accurately used [RQ2.2]

In order to select a suitable machine learning methodology to assess the MDM use of the test takers, we considered several algorithms that were capable of producing word embeddings. These included Recurrent Neural Networks (RNNs), Long-Short Term Memory (LSTMs) (Hochreiter & Schmiduber, 1997), and Transformers (Vaswani et al., 2017). Given their success in various downstream natural language processing (NLP) tasks in the literature, Transformers were chosen for this task. Additionally, they offer vastly reduced training times due to its ability to process entire sequences in parallel, through the use of 'attention mechanisms' that allows for tracking the relations between words across long text sequences in both forward and backward directions simultaneously.

The following classifiers/classification algorithms were selected to evaluate the performance of the appropriacy of MDM use by test takers:

- AdaBoost (Freund & Schapire, 1999)
- Decision Tree
- k-nearest neighbors classifier (kNN) (Zhang, Z., 2016)
- Multi-layer Perceptron (MLP)
- Naive Bayes (Zhang, 2004)
- Quadratic Discriminant Analysis (QDA)
- Random Forest Classifier (RFC) (Breiman, 2001)

(2) Embeddings

Each script was passed through a given embedding method in order to obtain word embeddings for all words contained in that script. For example, a Transformer mathematically encodes the words in context in the labelled dataset. A word is expressed in the form of a vector input (i.e., a string of numbers) which is called a word embedding.

A simple binary classifier requires a vector input for each data point (i.e., a string of numbers representing a given word in our dataset) in order to predict an output class. A vector representation of our textual data must be derived.

A given word in a sentence taken in isolation has little interpretability. The meaning of a word is dependent on its context and, as such, we must be able to encode information about a sequence of words in a single vector. Word embeddings give us a way to represent each word as an individual vector, whilst maintaining varying levels of contextual information in each embedding.

The majority of NLP tasks use Transformers to obtain these embeddings, given its state-of-the-art performance (SOTA) on benchmark NLP tasks as well as faster training times than conventional machine learning methods designed for sequential data, such as RNNs and LSTMs. Increasingly larger datasets are being used for training which has given rise to generalizable pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). The application of a pre-trained Transformer enabled us to make use of a model that has been trained on very large

datasets compared to the size of the dataset used in this report. As a result, the models provided a bootstrap mechanism for the work in this report.

(3) Automated classification by classifiers

The word embeddings were then used to perform the binary classification tasks, using the labelled dataset in order to train the classifier. Word embeddings serve as features that allow a classifier to group words with similar properties together. The classifier outputs 1 or 0 for each word (as coded in the labelled dataset).

The initial automated classification shows that in our labelled dataset, only 13% of words were labelled by the classifiers as MDMs (RQ2.1) and of these only 5.9% are labelled as not appropriately used (RQ2.2).

(4) Improvement of Classifiers

Based on the results of (3), measures were used to improve the performance of the classifiers. Any given algorithm has a number of parameters affecting the way it is able to learn from data, often significantly affecting classifier performance. To refine classifier performance, we also performed two fine-tuning measures:

a. Resampling methods are usually used to alter the composition of the dataset used for training such that the percentage of data belonging to each class is closer to 50%, generally improving classifier performance. Both undersampling of the majority class (the most frequently occurring class) and oversampling of the minority class (the least frequently occurring class) were trialed to observe the effects of class imbalance on the classifier. SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), SMOTEENN (SMOTE combined with edited-nearest-neighbours) and SMOTETomek (SMOTE combined with the use of Tomek Links) are resampling methods that have been used to create the resampled datasets.

b. Fine-tuning studies were conducted to find optimal learning parameters for our classifiers.

Results

RQ1: The Use of MDMs at Different CEFR Levels

Overall use of MDMs

The summary of human coding is presented in the form of descriptive statistics in Table 3, showing the ratio of scripts where at least one MDM (irrespective of accuracy) in each category appeared. Figure 2 presents the same information visually.

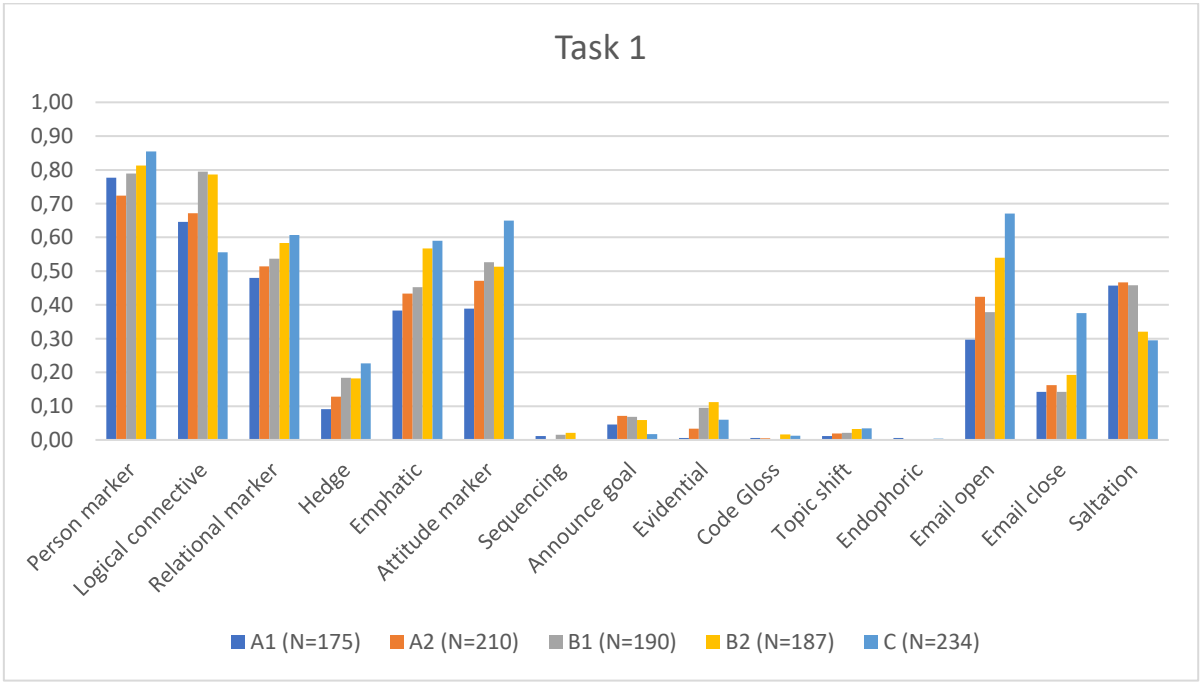
Table 3

Ratio of scripts with at least 1 MDM use (Task 1)

	A1 (N=175)	A2 (N=210)	B1 (N=190)	B2 (N=187)	C (N=234)	Whole (N=996)
Person marker	0.78	0.72	0.79	0.81	0.85	0.79
Logical connective	0.65	0.67	0.79	0.79	0.56	0.68
Relational marker	0.48	0.51	0.54	0.58	0.61	0.55
Hedge	0.09	0.13	0.18	0.18	0.23	0.17
Emphatic	0.38	0.43	0.45	0.57	0.59	0.49
Attitude marker	0.39	0.47	0.53	0.51	0.65	0.52
Sequencing	0.01	0.00	0.02	0.02	0.00	0.01
Announce goal	0.05	0.07	0.07	0.06	0.02	0.05
Evidential	0.01	0.03	0.09	0.11	0.06	0.06
Code Gloss	0.01	0.00	0.00	0.02	0.01	0.01
Topic shift	0.01	0.02	0.02	0.03	0.03	0.02
Endophoric	0.01	0.00	0.00	0.00	0.00	0.00
Email open	0.30	0.42	0.38	0.54	0.67	0.47
Email close	0.14	0.16	0.14	0.19	0.38	0.21
Saltation	0.46	0.47	0.46	0.32	0.29	0.40

Figure 2

Ratio of scripts with at least 1 MDM use (Task 1)



From Table 3 (and Figure 2), Task 1 (i.e., an informal email to a friend) elicited five interpersonal MDM groups (i.e., person marker, relational marker, hedge, emphatic and attitude marker), one textual MDM (i.e., logical connective) and three genre-specific MDM groups (i.e., saltation, email open and email close). The ratio of scripts that had at least one MDM in these categories tended to increase as the levels went up, except for logical connective and saltation marker. Specifically, the ratio of scripts containing

at least one logical connective in C scripts (0.56) were lower than B1 and B2 scripts (both 0.79), and for saltation markers, the ratio was lower at B2 (0.32) and C levels (0.29) than the B1 and below (0.46 and 0.47). This echoes with Carlsen (2010)'s finding that higher-level writers tend to rely less on logical connectives to establish discourse structure. Some MDM categories were hardly used in Task 1; namely, sequencing, announce goal, evidential, code gloss, topic shift, and endophoric MDMs. Different from the general perception that discourse competence develops at higher-proficiency levels, most of the lowest-proficiency A1 and A2 writers in this study used person markers and logical connectives, half used relational markers and saltation, and over one-third used emphatic and attitude markers.

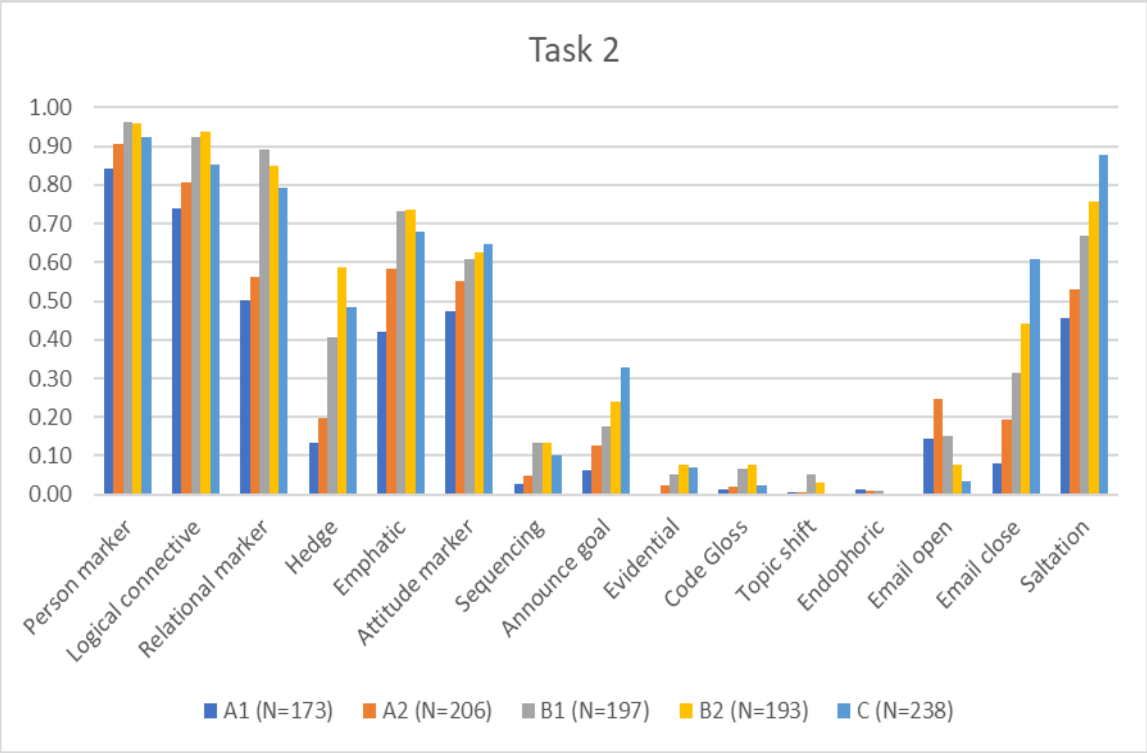
Table 4

Ratio of scripts with at least 1 MDM use (Task 2)

	A1 (N=173)	A2 (N=206)	B1 (N=197)	B2 (N=193)	C (N=238)	Whole (N=1007)
Person marker	0.84	0.91	0.96	0.96	0.92	0.92
Logical connective	0.74	0.81	0.92	0.94	0.85	0.85
Relational marker	0.50	0.56	0.89	0.85	0.79	0.73
Hedge	0.13	0.20	0.41	0.59	0.48	0.37
Emphatic	0.42	0.58	0.73	0.74	0.68	0.64
Attitude marker	0.47	0.55	0.61	0.63	0.65	0.59
Sequencing	0.03	0.05	0.13	0.13	0.10	0.09
Announce goal	0.06	0.13	0.18	0.24	0.33	0.19
Evidential	0.00	0.02	0.05	0.08	0.07	0.05
Code Gloss	0.01	0.02	0.07	0.08	0.03	0.04
Topic shift	0.01	0.00	0.05	0.03	0.00	0.02
Endophoric	0.01	0.01	0.01	0.00	0.00	0.01
Email open	0.14	0.25	0.15	0.08	0.03	0.13
Email close	0.08	0.19	0.31	0.44	0.61	0.34
Saltation	0.46	0.53	0.67	0.76	0.88	0.67

Figure 3

Ratio of scripts with at least 1 MDM use (Task 2)



In comparison to Task 1, Task 2 (i.e., a formal complaint email to a manager) elicited a wider range of MDM groups (see Table 4 and Figure 3). Test takers used five interpersonal MDM groups (i.e., person marker. Relational marker, hedge, emphatic and attitude marker), three textual MDM (i.e., logical connective, sequencing and announce goal) and three genre-specific MDM groups (i.e., saltation, email open and email close). On Task 2, it was not always C level candidates who revealed the highest ratio, although it is generally observed that more candidates use the MDMs at higher levels. It is notable that C level candidates produced more MDMs for saltation (0.88), to announce goals (0.33), and to close the email message (0.61) than the candidates at lower levels. This suggests that C level candidates might be more aware of the structure of a formal complaint email, in which they addressed and stated more clearly why they are writing to the person of power (e.g., school manager) while expressing their feelings (e.g., I am deeply disappointed that…) as well as closing the email often asking for a prompt response. The lowest-proficiency A1 and A2 writers, again, showed evidence of discourse competence through use of MDMs. A vast majority of A1 and A2 writers used person markers and logical connectives, half used relational, emphatic, attitude markers and saltation, and 20% of A2 writers used hedges. It is worth noting their different choices of MDMs between the two tasks, even though the difference was more subtle than that shown by the higher-level writers.

Accurate use of MDM at different CEFR Levels

Table 5 and Figure 4 present the average ratio of accurately used MDMs at different CEFR levels for Task 1. Table 6 and Figure 5 are for those for Task 2. It is clear that, in both tasks, the ratios of accurately used MDM are very similar across the CEFR levels—around 0.90—for most types of MDM. This means that when MDMs were used, candidates used them accurately regardless of their proficiency levels. The exceptions are the slightly lower ratios for announcing goals and email closing for both tasks. While, as aforementioned, lower-proficiency writers used these makers, higher-proficiency writers were more able to use them accurately. In comparison to the other used MDM categories (such as person markers

and relational markers), there are multiple ways to achieving announcing goals and email closing and often involve more than a single word. We can also see 'jagged' ratios for sequencing and endophoric MDMs in Task 1, but given the very small number of cases in these MDM (as shown in Table 5), this may not be a representative picture.

Table 5

Average ratio of accurately used MDM across CEFR levels (Task 1)

	A1 (N=175)	A2 (N=210)	B1 (N=190)	B2 (N=187)	C (N=234)
Person marker	0.97	0.98	0.99	1.00	1.00
Logical connective	0.99	0.99	0.99	0.99	0.99
Relational marker	0.98	0.98	0.99	0.98	1.00
Hedge	0.84	0.89	0.97	0.99	0.98
Emphatic	0.98	0.97	0.99	0.98	0.99
Attitude marker	0.90	0.97	0.97	0.99	1.00
Sequencing	0.50	-	1.00	0.75	-
Announce goal	0.67	0.65	0.70	0.72	0.78
Evidential	0.00	0.71	1.00	0.90	0.93
Code Gloss	1.00	1.00	-	1.00	1.00
Topic shift	1.00	1.00	1.00	1.00	1.00
Endophoric	1.00	-	-	-	1.00
Email open	0.96	0.93	0.92	0.96	0.96
Email close	0.75	0.62	0.49	0.60	0.76
Saltation	0.97	0.98	1.00	1.00	1.00

Figure 4

Average ratio of accurately used MDM (Task 1)

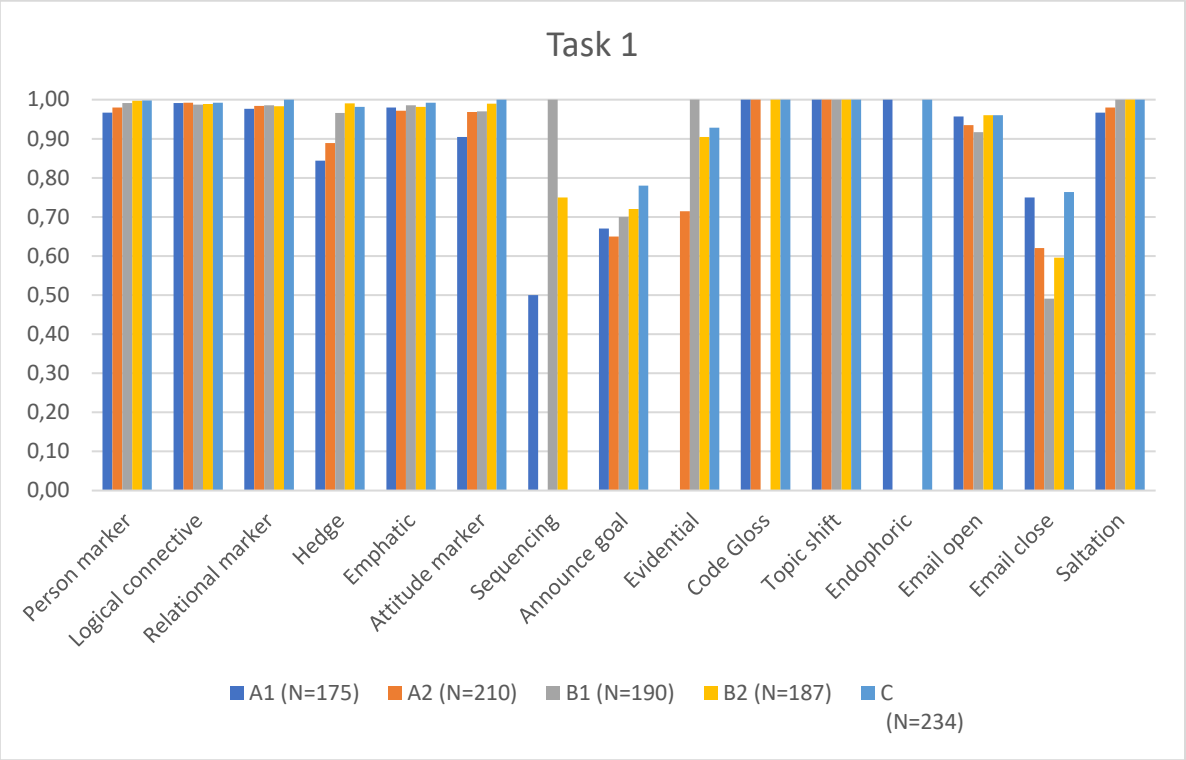


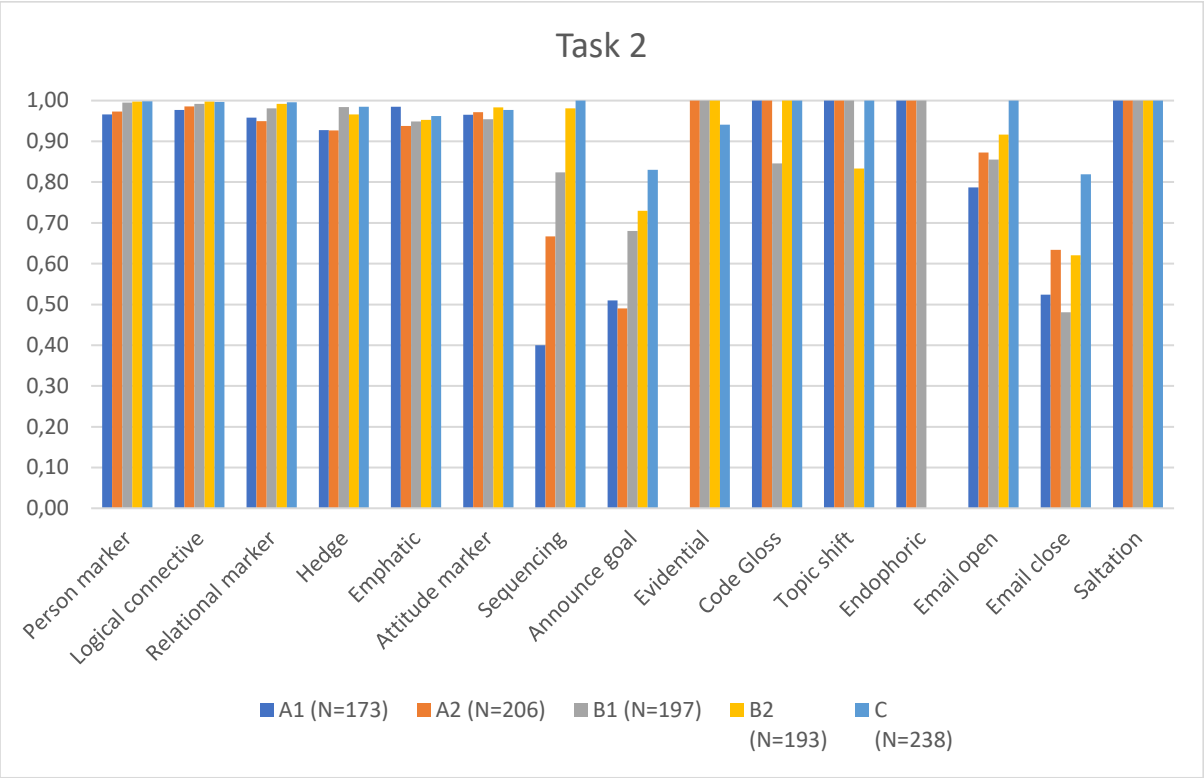
Table 6

Average ratio of accurately used MDM across CEFR levels (Task 2)

	A1 (N=173)	A2 (N=206)	B1 (N=197)	B2 (N=193)	C (N=238)
Person marker	0.97	0.97	1.00	1.00	1.00
Logical connective	0.98	0.99	0.99	1.00	1.00
Relational marker	0.96	0.95	0.98	0.99	1.00
Hedge	0.93	0.93	0.98	0.97	0.98
Emphatic	0.99	0.94	0.95	0.95	0.96
Attitude marker	0.97	0.97	0.95	0.98	0.98
Sequencing	0.40	0.67	0.82	0.98	1.00
Announce goal	0.51	0.49	0.68	0.73	0.83
Evidential	-	1.00	1.00	1.00	0.94
Code Gloss	1.00	1.00	0.85	1.00	1.00
Topic shift	1.00	1.00	1.00	0.83	1.00
Endophoric	1.00	1.00	1.00	-	-
Email open	0.79	0.87	0.86	0.92	1.00
Email close	0.52	0.63	0.48	0.62	0.82
Saltation	1.00	1.00	1.00	1.00	1.00

Figure 5

Average ratio of accurately used MDM (Task 2)



Kruskal-Wallis tests were then conducted to identify the differences in the ratio of accurately used MDMs (see Table 7). The results show significant differences in the accuracy of MDM use. Post-hoc pairwise comparisons identified some combinations of CEFR levels in which significant differences were found. However, the levels tended to be far apart, such as between A1 (beginner) and C (proficient learner).

Table 7

Results of Kruskal-Wallis tests (Task 1)

MDM Type	N	H	df	Sig.		Significant differences found between
Person marker	790	27.168	4	0.000	**	A1 and B2 (mean rank difference = 41.984, SE = 10.472, adjusted p=.001) A1 and C (mean rank difference = 45.640, SE = 9.86, adjusted p =.000) A2 and C (mean rank difference = 27.708, SE = 9.547, adjusted p=.037)
Logical connective	682	0.47	4	0.976		
Relational marker	545	5.119	4	0.275		
Hedge	165	8.08	4	0.089		
Emphatic	488	1.107	4	0.893		
Attitude marker	515	18.061	4	0.001	**	A1 and B2 (mean rank difference = 24.046, SE = 7.504, adjusted p=.014) A1 and C (mean rank difference = 26.765, SE = 6.907, adjusted p=.001)

Table 7 (Continued)

Results of Kruskal-Wallis tests (Task 1)

MDM Type	N	H	df	Sig.		Significant differences found between
Sequencing	9	1.571	2	0.456		
Announce goal	51	0.973	4	0.914		
Evidential	61	13.788	4	0.008	**	A1 and B1 (mean rank difference = 44.703, SE = 15.747, adjusted p=.045) B1 and C (mean rank difference = 44.525, SE = 12.481, adjusted p=.004)
Code Gloss	8	0	3	1.000		
Topic shift	24	0	4	1.000		
Endophoric	2	0	1	1.000		
Email open	471	3.938	4	0.414		
Email close	210	17.565	4	0.002	**	A1 and B2 (mean rank difference = 41.984, SE = 10.472, adjusted p=.001) A1 and C (mean rank difference = 45.640, SE = 9.86, adjusted p =.000) A2 and C (mean rank difference = 27.708, SE = 9.547, adjusted p=.037)
Saltation	394	7.14	4	0.129		

Table 8 presents the results of Kruskal-Wallis tests for Task 2. Like Task 1, the accuracy of use was found to be significantly different in some MDMs. The post-hoc pairwise comparisons identified differences between CEFR levels that are closer for some MDMs (e.g., B2 and C in email closing) in Task 2 than in Task 1. This is probably due to the nature of the email tasks as the MDMs used in formal emails (I'm writing to) tend to be more formulaic than those used in informal emails (e.g., do you know that?, I want to tell you …). The variation between the two tasks will be addressed again in RQ2.

Table 8

Results of Kruskal-Wallis tests (Task 2)

MDM Type	N	H	df	Sig.		Significant differences found between
Person marker	928	37.473	4	0.000	**	A1 and B1 (mean rank difference = 38.137, SE = 13.236, adjusted p=.000) A1 and B2 (mean rank difference = 49.816, SE = 13.314, adjusted p =.000) A1 and C (mean rank difference = 55.994, SE = 12.838, adjusted p =.000) A2 and B1 (mean rank difference = 38.913, SE = 12.389, adjusted p=..002) A2 and B2 (mean rank difference = 50.636, SE = 12.471, adjusted p=.000) A2 and C (mean rank difference = 56.769, SE = 11.962, adjusted p=.000)
Logical connective	860	8.018	4	0.091		
Relational marker	732	19.111	4	0.001	**	A2 and B2 (mean rank difference = 24.046, SE = 7.504, adjusted p=.014) A2 and C (mean rank difference = 26.765, SE = 6.907, adjusted p=.001)

Table 8 (Continued)

Results of Kruskal-Wallis tests (Task 2)

MDM Type	N	H	df	Sig.		Significant differences found between
Hedge	372	4.58	4	0.333		
Emphatic	641	4.104	4	0.392		
Attitude marker	591	5.096	4	0.278		
Sequencing	91	21.206	4	0.000	**	A1 and B2 (mean rank difference = 26.885, SE = 8.085, adjusted p=.009) A1 and C (mean rank difference = 28.500, SE = 8.139, adjusted p=.005) A2 and C (mean rank difference = 18.300, SE = 6.231, adjusted p=.033)
Announce goal	196	18.204	4	0.001	**	A1 and B2 (mean rank difference = 71.870, SE = 18.292, adjusted p=.001) A1 and C (mean rank difference =59.288, SE = 17.553, adjusted p=.007)
Evidential	47	1.765	3	0.623		
Code Gloss	40	4.263	4	0.372		
Topic shift	19	2.167	4	0.705		
Endophoric	6	0	2	1.000		
Email open	129	5.432	4	0.246		
Email close	346	56.427	4	0.000	**	A1 and C (mean rank difference = 75.880, SE = 25.959, adjusted p=.035) A2 and C (mean rank difference = 49.578, SE = 16.566, adjusted p =.028) B1 and B2 (mean rank difference = 42.128, SE = 15.491, adjusted p=.065) B1 and C (mean rank difference = 98.182, SE = 14.075, adjusted p=.000) B2 and C (mean rank difference = 56.054, SE = 12.671, adjusted p=.000)
Saltation	675	0	4	1.000		

We have so far reported the results of RQ1 regarding the use and accuracy of MDMs between the informal and formal email tasks by the L2 writers across the proficiency spectrum. In the next section and onwards, we present the results of RQ2 regarding the extent to which outcomes of RQ1 can be used to build AI models in relation to the classifier performance for whether an MDM or Not, the classifier performance for accurately or inaccurately used MDM, and the impact of task dependency.

RQ2: Classifications using a transformer-based AI Model

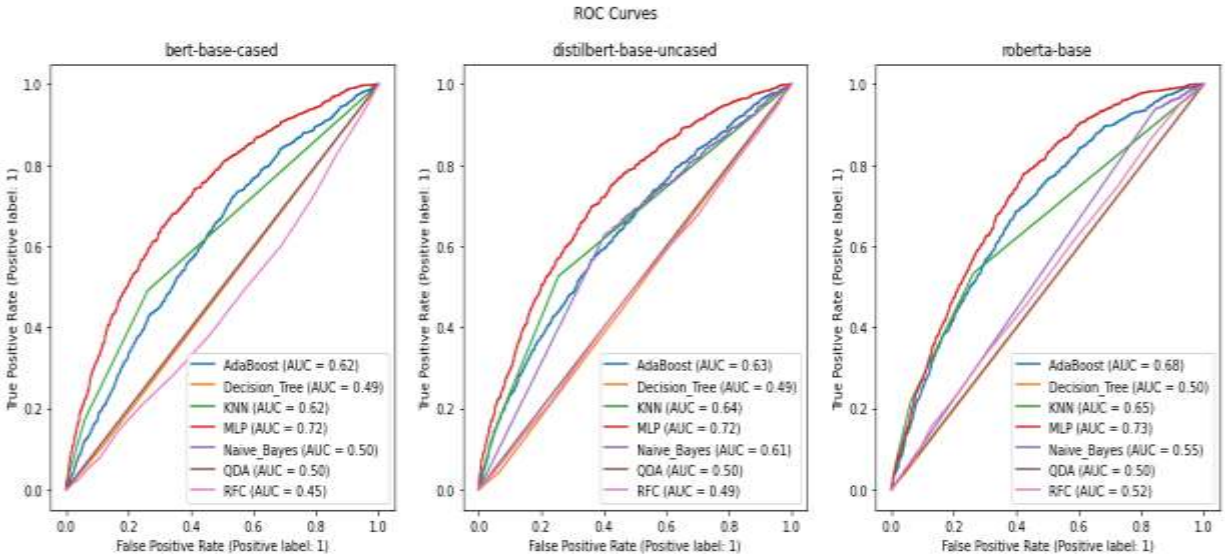
Word Embeddings

For all experimentation, our dataset was split into a train, validation and test set with 60%, 20% and 20% of the dataset belonging to each set respectively.

After consideration of benchmark performance and training times, the performance of the three shortlisted transformers BERT, RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) were evaluated and compared. An initial 10% of the overall dataset was used in order to reduce training times at this stage. Multiple out-of-the-box classifiers (i.e., using default learning parameters) were used together with the transformer architectures, to avoid the need to fine-tune classifier parameters. The resulting receiver operating characteristic (ROC) Curves are shown in Figure 6.

Figure 6

ROC curves of the 10% dataset of different transformer architectures evaluated on a range of out-of-the-box classifiers. Area under curve (AUC) scores are also shown



The distribution of resulting AUC scores[2] were relatively similar across the embedding methods. Among them, the MLP and the boosting algorithm, AdaBoost, showed the best out-of-the-box performance (ROC curves with points closer to the upper left of the graph show better performance due to their lower False Positive Rate for a given True Positive Rate).

Due to the limited variation in performance between embedding methods, BERT embeddings (far left in Figure 6) were selected for use. Owing to faster training times, a variation of the boosting algorithm, LightGBM (Ke et al., 2017), was used to evaluate performance. From this point forward, the entirety of the labelled dataset was used for experimentation unless otherwise specified.

Classifier Performance for Whether an MDM or Not

When evaluated on the 20% test set, with the other lines on the graph showing either classifiers trained on resampled datasets or fine-tuned classifiers. From the Precision-Recall curve[3] and the ROC, we can see an apparent trade-off between the opposing classes as the classification threshold is varied. However, even with the introduction of both under-sampling and oversampling techniques (i.e., ADASYN and SMOTE), we see limited changes in the metrics.

[2] AUC stands for "Area under the ROC Curve", which measures the area underneath the ROC curve from (0,0) to (1,1). The AUC score can also be thought of as the probability that a randomly chosen positively labelled prediction ranks higher than a negatively labelled prediction. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR).

[3] According to Shafi (2022), precision-recall is a useful measure of success of prediction when the classes are very imbalanced. Precision is calculated by dividing the true positives by anything that was predicted as a positive. Recall (or True Positive Rate) is calculated by dividing the true positives by anything that should have been predicted as positive. The precision-recall curve shows the trade-off between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

We now describe the evaluation of classifier performance for RQ2.1. The blue lines (LGB) in Figure 7 show the performance of the trained LightGBM classifier on the original dataset.

Figure 7

Precision-Recall curves and ROC curves of the baseline classifier, classifiers trained on resampled datasets and fine-tuned classifiers. Average precision (AP) and AUC scores are shown.

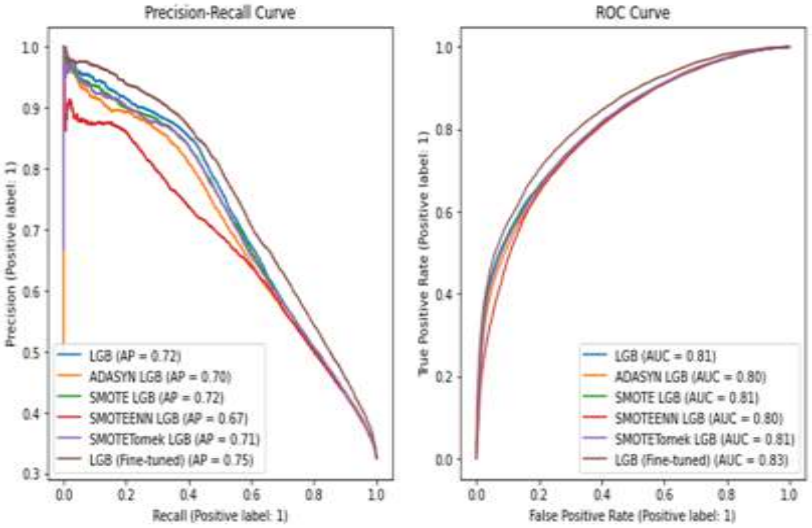


Table 9 shows the metrics for evaluating the accuracy of predictions by different resampling methods (i.e. Precision for class 0 and 1 (Pre0 and Pre1), Recall for class 0 and 1 (Rec0, Rec1), G-Mean and F1-Macro). For all the values, the closer to 1, the better the predictions are.

The addition of fine-tuning showed significant improvements in the recall of class 1, however, this came at the cost of a lower precision. According to the classification probability histogram of class 1 in Figure 8 between our baseline classifier and our fine-tuned model, LGB (Fine-tuned), we can see a definitive shift in the overlap between the classification of our two opposing classes. In the baseline model, class 1 exhibited a bimodal distribution. In our fine-tuned model, we see a rise in the bias of the classifier towards class 1, which results in higher confidence of correctly identifying a MDM at the cost of a rise in data from class 0 having a higher probability of being predicted as belonging to class 1 (as demonstrated by the larger tail of the blue line in the fine-tuned model). Whilst the confidence of our model has improved through fine-tuning, a tradeoff still exists and there is room for improvement.

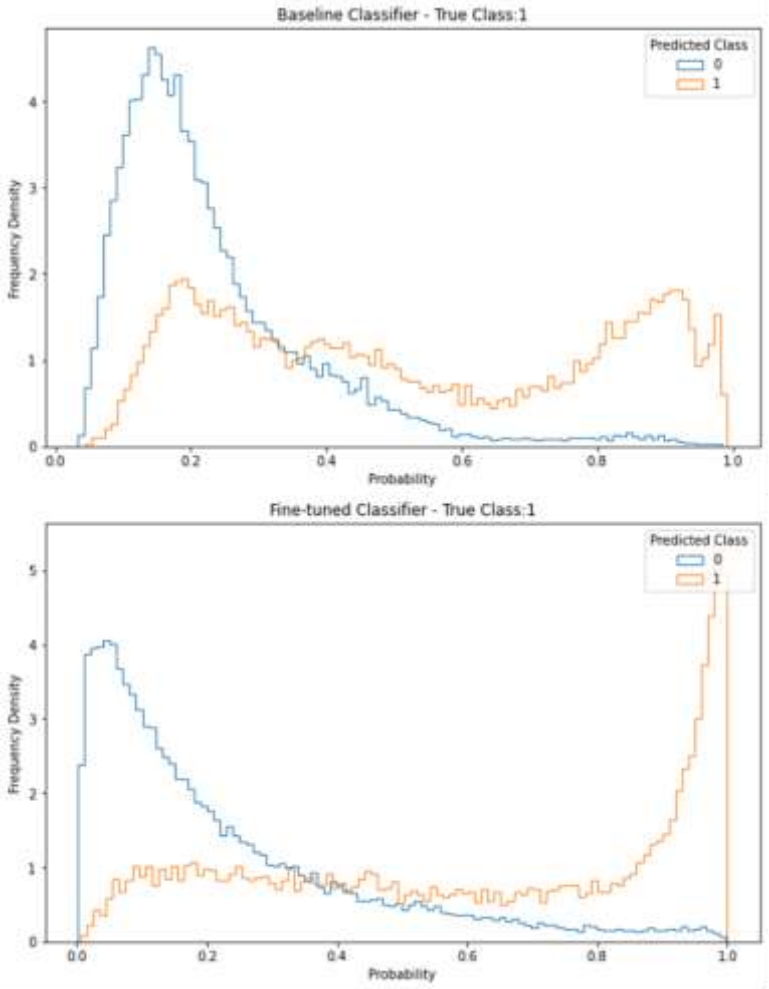
Table 9

Metrics for evaluating the accuracy of predictions (RQ2.1)

Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.79	0.79	0.94	0.48	0.67	0.73
ADASYN LGB	0.83	0.60	0.79	0.66	0.72	0.72
SMOTE LGB	0.82	0.63	0.82	0.64	0.72	0.73
SMOTEENN LGB	0.88	0.47	0.55	0.84	0.68	0.64
SMOTETomek LGB	0.82	0.63	0.83	0.63	0.72	0.73
LGB (Fine-tuned)	0.82	0.70	0.88	0.61	0.73	0.75

Figure 8

Frequency density of the probability of prediction of the baseline and fine-tuned classifier for class 1

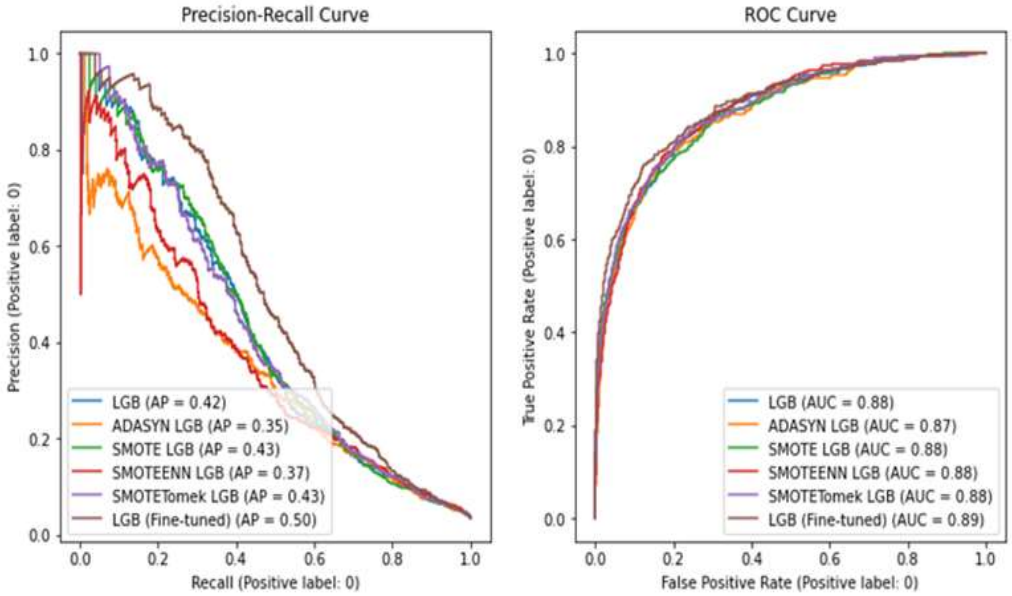


Classifier Performance for Accurately or Inaccurately Used MDM

This section describes the evaluation of classifier performance for RQ2.2. Once again, the blue lines in Figure 9 shows the performance of the trained LightGBM classifier on the original dataset, with the other lines on the graph involving LGB classifiers trained on resampled datasets or otherwise represent fine-tuned classifiers.

Figure 9

Precision-Recall curves and ROC curves of the baseline classifier, resampled datasets and fine-tuned classifiers. Average precision (AP) and AUC scores are shown.



From Table 10, we see all classifiers performing exceedingly well in predicting accurate MDM (class: 1), however they had very little success in confidently predicting inaccurate use cases (class: 0). The introduction of resampling techniques increased the recall of the classifier at heavy cost to the precision when compared with the baseline classifier. This trade-off is made even more apparent when looking at the Precision-Recall curve and AP scores shown in Figure 8, with an average AP score 0.40 amongst all classifiers.

Table 10

Metrics for evaluating the accuracy of predictions (RQ2.2)

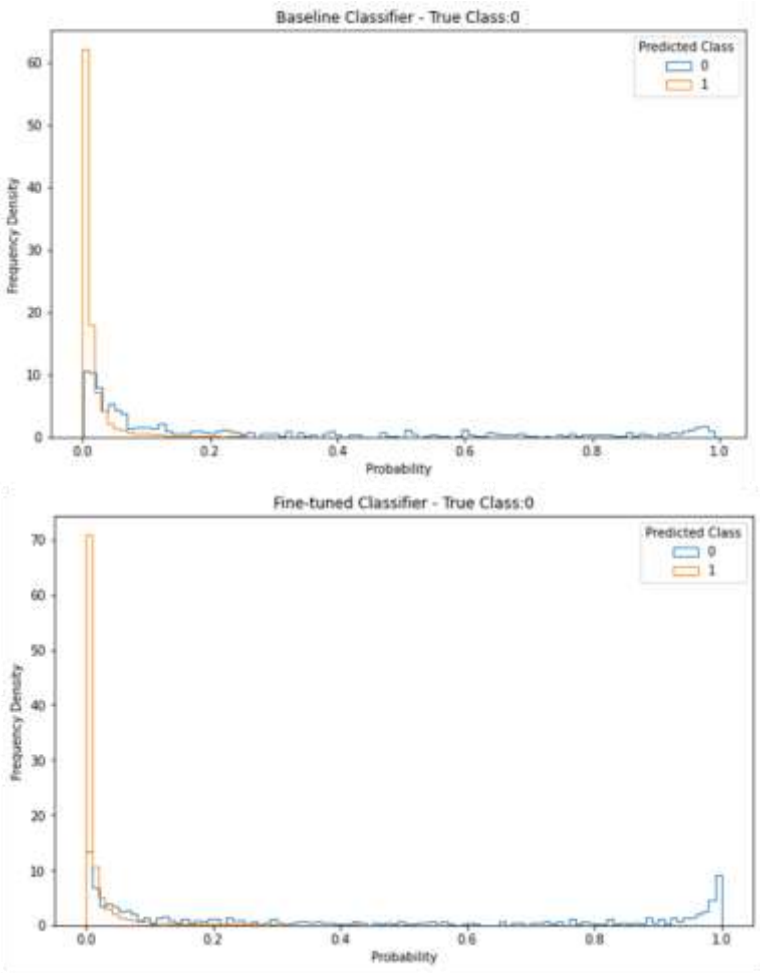
Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.74	0.24	0.97	1.00	0.49	0.67
ADASYN LGB	0.27	0.56	0.98	0.94	0.73	0.66
SMOTE LGB	0.29	0.55	0.98	0.95	0.72	0.67
SMOTEENN LGB	0.19	0.68	0.99	0.89	0.78	0.61
SMOTETomek LGB	0.30	0.55	0.98	0.95	0.72	0.68
LGB (Fine-tuned)	0.67	0.38	0.98	0.99	0.62	0.74

Furthermore, looking at the classification probability histograms for class 0 shown in Figure 7, we see improved performance in the discriminative ability of our fine-tuned classifier, LGB (Fine-tuned), when compared to baseline performance. However, a significant portion of our inaccurate MDM use cases were predicted as having a very low probability of belonging to class 0. Due to the aforementioned overwhelming imbalance in our dataset (roughly 24 accurate use cases for every inaccurate use case), the consequences of adjusting the classification threshold were significant. Whilst only a small percentage of either class was affected by changes to the classification threshold, the class imbalance resulted in a much larger absolute value of accurate use cases being misclassified as the threshold

decreases. As such, attempting to include these low confidence occurrences is not feasible and our classifier, as a result, is only able to predict a portion of inaccurate use cases well.

Figure 10

Frequency density of the probability of prediction of the baseline and fine-tuned classifier for class 0



Task Dependency

In addition, we tested the classifiers dependence on a given task by training classifiers on one task exclusively, whilst using data from the other task to test its performance. We show results from both research questions RQ2.1 and RQ2.2 on classifiers trained with all the training split data (the baseline classifier) alongside classifiers trained solely on either Task 1 or Task 2 training data whilst using the unused task for testing.

For RQ2.1, our results from Figure 11 and Table 11 show that a classifier trained solely on Task 2 data, classifier 2, is better capable of generalizing on an unseen task than a classifier trained solely on Task 1 data, classifier 1. Classifier 2 outperforms the baseline classifier in several areas as shown by our established evaluation metrics.

Figure 11

Precision-Recall curves and ROC curves of the baseline classifier and classifiers trained and tested on differing classes

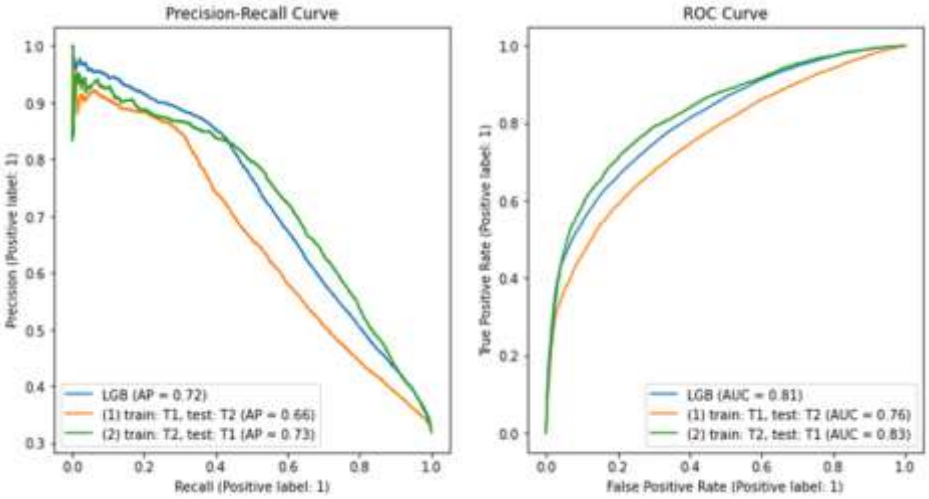


Table 11

Metrics for evaluating the dependence of the classifier on a given task (RQ2.1)

Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.79	0.79	0.94	0.48	0.67	0.73
(1) train: T1, test: T2	0.77	0.71	0.91	0.44	0.63	0.69
(2) train: T2, test: T1	0.81	0.76	0.92	0.55	0.71	0.75

However, for RQ2.2, the inverse is true; with classifier 1 giving better performance than classifier 2, as shown in Figure 12 and Table 12. Classifier 2 severely underperforms in all areas pertinent to the classification of class 0 when compared with our baseline and classifier 1.

Figure 12

Precision-Recall curves and ROC curves of the baseline classifier and classifiers trained and tested on differing classes

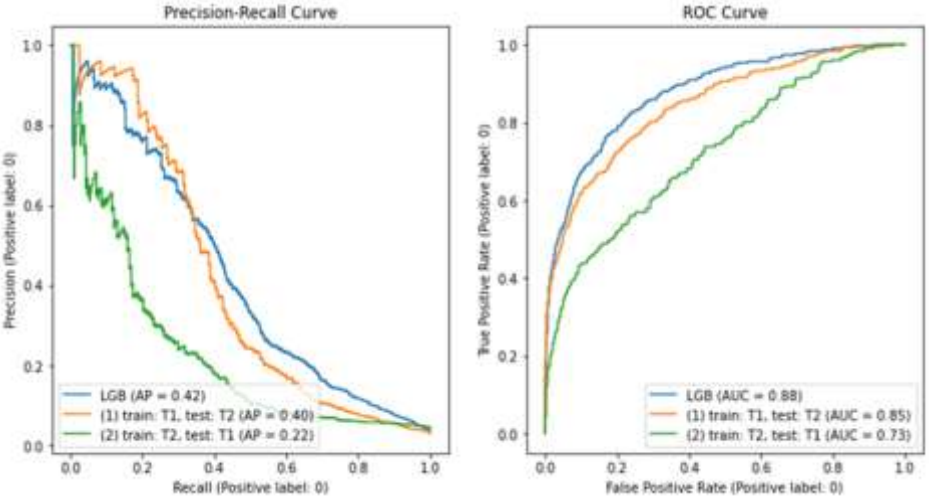


Table 13

Metrics for evaluating the dependence of the classifier on a given task (RQ2.2)

Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.74	0.24	0.97	1.00	0.49	0.67
train: Task 1, test: Task 2	0.80	0.23	0.98	1.00	0.48	0.67
train: Task 2, test: Task 1	0.61	0.10	0.96	1.00	0.31	0.57

To summarise, the detection of MDMs (RQ2.1) is much less dependent on a given task than detecting the appropriacy of use (RQ2.2). However, with only two tasks to test dependency, a definitive conclusion cannot be drawn for either of our research questions as to whether our classifiers can be said to be task-agnostic.

Discussion

Questions and Limitations

Findings of RQ1 show that, different from the general notion that discourse competence emerges at the B2/C1 threshold, L2 writers start to display discourse competence (or at least in terms of the use of MDM) even at CEFR A1, A2 and B1 levels. However, C level candidates are clearly most aware of the difference between the informal and formal essay tasks and most able to adjust their use of MDM accordingly. Previous studies (e.g., Knoch et al., 2014) found that lower-level writers used proportionally more MDMs overall than more proficient writers, or that more proficient writers used a significantly wider range of MDMs than lower-level writers (e.g., Bax et al., 2019). However, this is not the case in the current study. This indicates the importance of considering genre variation in research of MDMs (or discourse competence) in L2 writing. For example, the task investigated in Knoch et al., (2004) was an essay task whereas informal and formal emails were investigated in the current study. Emails are served as communication tools whereas essays are often used to displace understanding of a certain topic and to introduce one's (new) perspectives. It can be argued that the main goal of emails is to communicate with a specific (usually known) audience, whereas essays are often expository and argumentative aiming to address a wider unknown audience. Discourse competence in email writing is essential to ensure effective communication and to conform to the real-world expectation of audience awareness and appropriateness. For example, as shown in this study, an informal email to a friend allows for a more relaxed, friendly style through the use of interpersonal markers (e.g., person marker, relational marker, emphatic an attitude marker). Textual and genre-specific MDMs are less important, especially in short informal emails. Formal emails, even shorter ones, required more careful thought on tone and structure. As demonstrated by the L2 writers in this study, this can be achieved by the use of interpersonal markers (e.g., hedge), textual markers (e.g., sequencing and achieving goals). Genre-specific MDMs in formal emails (e.g., saltation, email open and email close) are deemed necessary. The findings show that even the lowest-proficiency L2 writers demonstrated some evidence of discourse competence in email writing through the use of MDMs. Nevertheless, they may struggle to demonstrate it in essay writing where textual and genre-specific MDMs might play a larger role. Perhaps there needs to be a different framework for evaluating the accuracy of MDM use according to the genres and levels of formality required. This also indicates the benefits of evaluating discourse competence through a detailed analysis of the use of individual MDMs in different categories in order to capture the nuanced evidence of the development of discourse competence in L2 writers.

In terms of accuracy of use, the findings show that L2 writers across the CEFR levels seem to use most of the MDMs accurately, more so in the formal email when use of MDM was more formulaic than the informal email. Nevertheless, several MDM types appeared to be exceptions, showing an upward trend

as the levels went up: hedge and announce goals on Task 1 and sequencing, announce goals and email close on Task 2. These MDM categories tend to involve more than a single word and thus there is a higher chance for lower-proficiency writers to make mistakes. Higher-proficiency writers, on the other hand, can demonstrate their discourse competence through accurate use of these phrasal/clausal MDMs.

Nevertheless, the overall uniform pattern in terms of accuracy of MDM use led to extremely small numbers of inaccurate MDM use. This further led to imbalanced data distribution for automated classification which led to attempting various resampling methodologies in RQ2.1 and RQ2.2, since a heavily imbalanced binary classification task can lead to the classifiers exhibiting heavy bias to the majority class. The tested resampling methods all performed relatively similarly, increasing the precision of the minority class as well as the recall of the majority class. Fine-tuning of the basic classifier was also attempted to improve performance and, whilst not consistently resulting in the highest individual class precision or recall scores, resulted in the highest F1-macro score. One might wonder: would focusing on MDMs be useful at all for classifying (and predicting discourse competence in L2 writing further in the future) and overall proficiency levels of the candidates? The simple answer is that measures of MDM would not be very useful alone in building automated essay scoring systems; they need to be incorporated with other features.

Moreover, the results for RQ2.1 and 2.2 clearly showed that there is a tradeoff between the classification and misclassification of 1 and 0. Of course, the ideal scenario is that we do not have false positives and false negatives, but it was not possible to achieve with the current dataset. We therefore need to consider what should be prioritized by asking the following questions:

- Do we worry more about misclassifying accurate use cases as inaccurate or the other way around?
- Do we worry more about identifying MDMs at the tradeoff of misclassifying?

Would judging accurately used MDM to be inaccurate be more damaging to the candidate's scores? Or would judging inaccurately used MDM to be accurate be worse for candidates? Misclassifying accurate and inaccurate MDM use can have significant consequences for learner assessment. When accurate MDM usage is misclassified as inaccurate, learners may receive unfairly low scores, leading to a loss of confidence in their writing abilities. For instance, a learner who effectively uses transition markers like "however" or "therefore" might be penalized if the system misinterprets their usage as incorrect, discouraging them from experimenting with more advanced language structures. Conversely, misclassifying inaccurate MDM usage as accurate can result in inflated scores, giving learners a false sense of mastery. Ultimately, the relative importance of these two errors depends on the candidates' learning stage. For beginners, it may be crucial to minimize errors where correct usage is judged as incorrect, because this type of error can discourage those who are making genuine progress and may lead them to question their understanding of the writing skills they are developing. On the other hand, for more advanced learners, the focus should perhaps shift to identifying and addressing incorrect usage to help students achieve greater accuracy in their writing. With a larger number of data points for our minority class (inaccurate use cases), weighting techniques could be explored to minimize errors of a given type depending on the writer's CEFR level.

One issue to bear in mind is that, although MDMs are found vitally important in articles, journals and newspapers (e.g., Hyland, 2005; Dafouz-Milne, 2008), they may carry slightly different 'weight' in email messages–more personal, shorter pieces of writing that are addressed to one person. Therefore, in order to explore answers to this set of difficult trade-off questions, we will need to scrutinize the construct of discourse competence that is being measured by the commonly used email tasks in L2 writing assessments and how the ability to use MDM is considered to contribute to the construct. For example, would having a specific MDM accuracy framework for informal writing than formal genres be appropriate or viable? It would also be important to consider the balance between accuracy, complexity and appropriateness.

Another question is how different the results might be if we are to expand the coding scheme to tap into pragmatic appropriacy. As described in the methods section, we employed a dichotomous coding scheme in this feasibility study, which required making unambiguous judgements on the use of MDM,

narrowing the range of errors that we coded for. They were mostly grammatical errors that surround the MDM but considering that the main role of MDM is to involve the message recipients, signpost, and communicate the writer's stance effectively, grammatical accuracy may only play a small part in it. Coding for pragmatic appropriacy will require polytomous coding (for example, 3: 'appropriate', 2: 'acceptable', 1: 'not appropriate') in order to capture other important aspects of MDM use. Furthermore, it might contribute to differentiate better between tasks with recipients with differential social status (i.e. a friend/classmate or a school manager), which could not be achieved with the dichotomous coding scheme in this study. Since the two email tasks used in the current study are designed to tap into the ability to use language according to different situations and recipients, coding for pragmatic appropriacy appears to be the clear next step forward. This can form one of the additional features to incorporate in automated classification, although it will be a more resource-intensive study which requires a bigger labelled dataset.

Conclusion and Future Work

In summary, the contributions of this study are three-fold. Firstly, it offers valuable insights within the context of Explainable AI. Transparency and explainability in automated scoring models are critical for ensuring fairness and stakeholder understanding in language tests. By integrating MDM usage and accuracy into the scoring framework, this research moves beyond frequency-based evaluation. The finding that the range of MDM use and accuracy are highly task-dependent highlights the need for task-specific tuning rather than relying on generalized models, which contribute to the design of AI-based systems that are both practical and explainable in educational applications.

Moreover, given that MDM usage reflects discourse competence, this study also makes significant contributions to the current understanding of L2 writing development. While previous research has often underestimated the discourse competence of lower-proficiency learners, this study demonstrates that even these learners exhibit evidence of discourse competence through their accurate use of MDMs as well as their choice of MDMs in response to genre. This finding suggests that L2 learners may develop aspects of discourse competence earlier than traditionally assumed, which offers a new perspective on how discourse analysis can be integrated as a core element in AEEs.

From the perspective of expanding the construct representation in automated scoring systems, this study provides a critical examination of the limitations of many AEE models, which have heavily relied on vocabulary and grammar features. By exploring the feasibility of incorporating MDMs as predictive features, this research demonstrates the potential for construct expansion. However, the task dependency of accuracy classification and the data imbalance—caused by the predominance of correct MDM usage—present challenges that need to be addressed.

There have been many AEE studies utilizing both handcrafted features as well as features obtained through the use of a deep neural network, however no definitive answer exists as to whether either will perform better on any given dataset. Several papers (e.g., Liu et al., 2019; Lin et al., 2020) suggest improved performance by training classifiers on a combination of both such feature types. One of the handcrafted features in the future could incorporate 'whether or not certain MDM are used' in the scripts. All the codes assigned in this study by the human coders indicated if MDM was used; their absence was not coded and therefore was not taken into account in automated classification. However, given accuracy use was consistently high across levels, it is possible that the absence of certain MDM could be good indicators that differentiate between the levels of proficiency. For example, the accurate use of MDM for announcing goals increased with the levels especially in the formal email task (RQ1). The absence of this MDM type might also differentiate effectively between CEFR levels, contributing to improving classifier performance. Another handcrafted feature could be formed using the genre of the text that we are attempting to analyze. Whilst transformer-based embeddings are context aware, explicitly introducing the genre of the text into the set of features fed into a classifier could allow for better performance over the detection of certain MDMs within said genres.

Another potential limitation of the current methodology surrounds the idea of phrasal MDMs. These are phrases consisting of words that in isolation are not MDMs, but when combined together are classed as an MDM. Currently, our transformer-based architecture returns a word embedding for each individual word that was input into it and these individual word embeddings are then classified. Whilst these embeddings are context dependent, labelling each individual word within that phrase as an MDM in its own right and training a binary classifier based on this data (RQ2.1) may be misleading and lead to the model underperforming when generalizing to the larger dataset. Future work could involve creating an extension of our word embedding methodology in which we form phrasal embeddings to represent our phrasal MDMs to assess the impact of this decision within our current framework. Phrasal embeddings could be formed using the word embeddings associated with words in the phrase we wish to obtain phrasal embeddings for (e.g., an average of all relevant word embeddings to obtain a singular phrasal embedding).

Finally, as previously mentioned, a dichotomous coding scheme was chosen for this particular study. Extending this coding scheme to a polytomous coding scheme may better allow a machine-learning model to discern the nuances associated with MDM usage. For example, as mentioned in Procedures for RQ1, less accurate MDM usage was not counted as inaccurate usage for our binary classification task. A polytomous coding scheme could allow a classifier to better distinguish such levels of MDM use since we currently have both less accurate use cases and extremely accurate use cases both having the same label. Additionally, this would somewhat reduce the class imbalance since data points would be moved from the current majority class (accurate use cases) to any intermediary labels included in the new coding scheme. However, it is still likely that collecting more data involving inaccurate use cases would allow the classifier to better establish class boundaries. Whilst the basis of the methodology would remain the same, using transformer-based architectures to extract word embeddings, our binary classification task could be extended to a multi-class classification task. Given the lack of research directly concerning MDMs in machine learning literature, we cannot definitively say how any of these factors would affect performance for our given tasks; however, all are valid areas that should be investigated in further work. Another area of research is how our results would impact skilled labour across various sectors, domains and contexts from construction, manufacturing up to medicine.

Declarations

Gen-AI use : The authors of this article declare (Declaration Form #: 212241332) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Ethical Approval: Ethical approval for the study was received from University of Bedfordshire, CRELLA Research Institute Ethics Committee on 14.04.2022.

Author Contribution: Sathena Chan: funding application, project lead, conceptualization, methodology, data analysis, writing, revision and editing. Manoranjan Sathyamurthy: methodology (RQ2), data analysis (RQ2), writing and revision. Chihiro Inoue: funding application, conceptualization, methodology (RQ1), data analysis (RQ1), writing, revision and supervision. Michael Bax: funding application, conceptualization, review and editing. Johnathan Jones: data analysis (RQ1). John Oyekan: funding application, conceptualization, methodology (RQ2), data analysis (RQ2), writing, revision and supervision.

Funding: The research was funded by the British Council Aptis Research Grants 2021.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no competing interests to disclose.

References

Adel, A. (2006). Metadiscourse in L1 and L2 English. John Benjamins Publishing. https://doi.org/10.1075/scl.24

Bachman, L. and Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Barkaoui, K. (2016). What changes and what doesn't? An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts. *IELTS Research Reports Online Series*, vol. 2016/3, 1–55.

Bax, S., D. Waller and Nakatsuhara, F. (2019). Researching L2 writers' use of MDM at intermediate and advanced levels, *System*, *83*, 79-95. https://doi.org/10.1016/j.system.2019.02.010

Breiman (2001). Random Forests, *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brezina V. & Gablasova, D. (2015) Is There a Core General Vocabulary? Introducing the *New General Service List*, *Applied Linguistics*, 36(1), 1-22, https://doi.org/10.1093/applin/amt018

Burneikaitė, N. (2008) "Metadiscourse in Linguistics Master's Theses in English L1 and L2", Kalbotyra, 59, pp. 38–47. doi:10.15388/Klbt.2008.7591.

Camiciottoli, B. C. (2003). Metadiscourse and ESP reading comprehension. *Reading In A Foreign Language*, *15*(1), 28–44. https://nflrc.hawaii.edu/rfl/item/69

Carlsen, C. (2010). Discourse connectives across CEFR-levels: A corpus based study. In I. Bartning, M. Maatin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and Language testing research* (pp. 191-210). European Second Language Association.

Chapelle, C. A. and Chung, Y-R. (2010). The Promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–315. https://doi.org/10.1177/0265532210364405

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321-357. https://doi.org/10.1613/jair.953

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press. https://doi.org/10.1017/CHOL9780521221283

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence, 14* (5), 771-780. http://www.yorku.ca/gisweb/eats4400/boost.pdf

Crompton, P. (2012). Characterising hedging in undergraduate essays by Middle-Eastern students. *Asian ESP Journal*, 8(2), 55-78. http://asian-esp-journal.com/wp-content/uploads/2013/11/Volume-8-2.pdf

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of deep Bidirectional Transformers for language understanding*. https://arxiv.org/pdf/1810.04805.pdf

Jarvis, S. (2013). 'Defining and measuring lexical diversity.' in S. Jarvis and M.H. Daller (eds.) *Vocabulary knowledge: human ratings and automated measures.* John Benjamins, pp. 13-44.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE. https//doi.org/10.1109/IJCNN.2008.4633969

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyland, K. (2005). *Metadiscourse*: Exploring Interaction in Writing. Bloomsbury Publishing.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems, 30* (NIPS 2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An Investigation of the Effect of Task Type on the Discourse Produced by Students at Various Score Levels in the TOEFL iBT ® Writing Test. *ETS Research Report Series*, *23*, 14–43. https://doi.org/10.1002/ets2.12038

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, *33*, 21-34. https://doi.org/10.1016/j.jslw.2016.06.004

Lialin, V., Deshpande, V. & Rumshisky, A. (2023). Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647. https://doi.org/10.48550/arXiv.2303.15647

Lin, W., Hasenstab, K., Chnha, G. M. & Schwartzman, A. (2020) *Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment*, *Scientific Reports, 10*, 20336. https://doi.org/10.1038/s41598-020-77264-y

Lin, T., Wang, Y., Liu, X. and Qiu, X. (2022). A survey of transformers. *AI open*, 3, 111-132. https://doi.org/10.48550/arXiv.2106.04554

Liu, J., Xu, Y. and Zhu, Y. (2019) 'Automated Essay Scoring based on Two-Stage Learning', *arXiv [cs.CL]*. https://doi.org/10.48550/arXiv.1901.07744

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

O'Loughlin, K. (2013). Investigating lexical validity in the Pearson Test of English Academic. Pearson Research Reports, p.1-21. https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/6iHE8HxuGJT3OMAgFoXxwV/88d7be6a2d43a9274d12fe7de838fef0/Investigating_lexical_validity_in_the_Pearson_Test_of_English_Academic-_Kieran_O___Loughlin.pdf

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., and Dunn, K. (2020). Aptis General Technical Manual Version 2.2. https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Owen, N., Shrestha, P. and Bax, S. (2021). Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Language (CEFR) levels assessed in the Aptis test. *ARAGs Research Reports Online, AR- G/2021/1*.

Sanford, S. (2012). *A comparison of metadiscourse markers and writing quality in adolescent written narratives.* Missoula: Unpublished MSc thesis. The University of Montana.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Schiffrin, D., Tannen, D., & Hamilton, H. (2001). *The handbook of discourse analysis.* Blackwell Publishers Ltd.

Zhang, H. (2004). The optimality of Naive Bayes. https://typeset.io/papers/the-optimality-of-naive-bayes-r4zge3fp91

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine, 4*(11), 218. https://doi.org/10.21037/atm.2016.03.37

Appendix

Appendix 1

Full list of MDMs analysed

Announce Goals (Frame marker)			
here I will	my purpose	the aim	I intend
I seek	I wish	I argue	I propose
I suggest	I discuss	I would like to	I will focus on
we will focus on	I will emphasise	we will emphasise	my goal is
in this section	in this chapter	here I do this	here I will
Code glosses			
put another way	for example	for instance	e.g.
i.e.	that is	that is to say	namely
in other words	this means	which means	in fact
Viz.	specifically	such as	
known as	defined as	called	
Endophorics			
see	noted	discussed below	discussed above
discussed earlier	discussed later	discussed before	section
chapter	fig	figure	table
example	page		
Hedges			
apparently	appear to be	approximately	assume
believed	certain extent	certain level	certain amount
could	couldn't	doubt	essentially
estimate	frequently	generally	in general
indicate	largely	likely	mainly
may	maybe	might	mostly
often	perhaps	plausible	possible
possibly	presumably	probable	probably
relatively	seems	sometimes	somewhat
suggest	suspect	unlikely	uncertain
unclear	usually	would	wouldn't
little	not understood	almost	
Logical connectives			
but	therefore	thereby	so
so as to	in addition	similarly	equally
likewise	moreover	furthermore	in contrast
by contrast	as a result	the result is	result in
since	because	consequently	as a consequence
accordingly	on the other hand	on the contrary	however
besides	also	whereas	while
although	even though	though	yet
nevertheless	nonetheless	hence	thus
leads to	or	and	
Relational markers			
incidentally	determine	consider	imagine
by the way	let us	let's	lets
let	notice	our	recall
note	us	we	you
our	one's	assume	think about
your			
Attitude markers			
admittedly	I agree	amazingly	unusually
accurately	correctly	curiously	disappointing
disagree	even	fortunately	have to