Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels

Stephen Bax [a], Fumiyo Nakatsuhara [a], and Daniel Waller [b]

[a] *University of Bedfordshire, Centre for Research in English Language Learning and Assessment (CRELLA), Room 118, Putteridge Bury Campus, Luton LU2 8LE, United Kingdom*

[b] *University of Central Lancashire, School of Language & Global Studies, University of Central Lancashire, Preston, PR1 2HE, United Kingdom*

Abstract

Metadiscourse markers refer to aspects of text organisation or indicate a writer's stance towards the text's content or towards the reader (Hyland, 2004:109). The CEFR (Council of Europe, 2001) indicates that one of the key areas of development anticipated between levels B2 and C1 is an increasing variety of discourse markers and growing acknowledgement of the intended audience by learners. This study represents the first large-scale project of the metadiscourse of general second language learner writing, through the analysis of 281 metadiscourse markers in 13 categories, from 900 exam scripts at CEFR B2-C2 levels. The study employed the online text analysis tool Text Inspector (Bax, 2012), in conjunction with human analysts. The findings revealed that higher level writers used fewer metadiscourse markers than lower level writers, but used a significantly wider range of 8 of the 13 classes of markers. The study also demonstrated the crucial importance of analysing not only the behaviour of whole classes of metadiscourse items but also the individual items themselves. The findings are of potential interest to those involved in the development of assessment scales at different levels of the CEFR, or to teachers interested in aiding the development of learners.

Highlights

- Advanced L2 writers use fewer metadiscourse markers than intermediate writers
- Advanced L2 writers can express text organisation without explicit markers
- Higher-level writers use a greater range of metadiscourse markers
- Higher-level writers abandon simple markers in favour of other makers

Keywords

metadiscourse markers; CEFR; L2 writing; language testing; Text Inspector; essay writing; text organisation; writer’s stance; online text analysis

Acknowledgements and Tribute

This paper reports on research funded through the Cambridge English Funded Research Programme, 2012-13. We would also like to acknowledge Professor Stephen Bax for the development of Text Inspector that allowed for the automated analysis of a large set of data. Crucially, Text Inspector also made it possible for the researchers to review the use of the individual markers in the context within which they occurred. This research has informed the refinements of the metadiscourse functionality of Text Inspector, and the flexibility of Text Inspector to incorporate fine adjustments based on research evidence is one of the strengths of the analysis tool. The development of Text Inspector is one of the many contributions that Professor Stephen Bax made to the field of language learning and assessment, and this paper is a tribute to his great work.

1. Introduction

Research into metadiscourse markers, defined by Hyland (2004:109) as “those aspects of the text which explicitly refer to the organisation of the discourse or the writer’s stance towards either its content or the reader”, has expanded considerably in recent years. This expansion is due to the growing awareness of the role of such markers in signalling important organisational and attitudinal dimensions in a variety of genres, particularly in investigations of academic genres. There still remains, however, a significant gap in the field, specifically with regard to the use of metadiscourse markers at different levels of second language (L2) writing proficiency. This question, of potential significance to millions of learners, educators, test developers and applied linguists worldwide, is therefore the focus of this research project, based on evidence drawn from a large-scale study of writing scripts assessed at Common European Framework of Reference (CEFR; Council of Europe, 2001) levels B2-C2.

2. Background

Discourse competence, long considered a key index of language learners’ developing proficiency, is seen by the CEFR as a core criterion for higher levels of language ability, with “a new focus on discourse skills” from Level B2 upwards (Council of Europe, 2001:35). In line with this focus on discourse at higher levels, CEFR descriptors at C1 and C2 levels explicitly reference discourse when specifying expected competences. In this light, analysis of discourse markers in L2 learners’ speech and writing could offer an important indicator for distinguishing criterially between language learners at different CEFR levels, especially upper levels (Weir, 2005). Shaw and Weir’s (2007:251) comment that this is a “very much under-researched area” still holds today, as does their view of the potential value of discourse features in contributing to "further grounding of distinctions between levels". Much of the prior work on discourse markers is based on studies using smaller data sets (e.g. Buneikaite, 2008; Intaraprawat & Steffensen, 1995; Noble, 2010) or focus on comparisons with native speakers in specific types of academic writing (e.g. Ädel, 2006; Crompton, 2012; Hyland, 2004; Lee & Deakin, 2016; Ozdemir & Longo, 2014). The present paper, however, is interested in large-scale data sets that compare different levels of learners. To date, no study has directly examined differential use of discourse markers in General English writing (as opposed to specific or academic English) on any significant scale at CEFR B2-C2 levels. This gap in the literature is the focus of this paper.

2.1. Metadiscourse markers in discourse

The analysis of metadiscourse markers is of potential value because of the key relationship between discourse competence and the observable linguistic features of a text. As McCarthy and Carter (1994:174) point out, "linguistic competence cannot be separated from discourse competence". Although early discussions of metadiscourse markers tended to downplay their role in texts (e.g. Harris, 1959), other research has acknowledged their significance and started to examine their role in organising message and intention, as well as in authorial comment on the content, as part of the wider communicative purpose of a text (e.g. Crismore & Farnsworth, 1990:119; Hyland, 2004:109; Vande Kopple, 1985:83)

When seeking evidence of an author's intentions in a text, it is important to look not only at the macro-level and features such as moves (Bhatia, 1993; Swales, 1990), but also at the micro-level, and in particular at specific markers within the text (Hyland & Tse, 2004:161). Metadiscourse markers are potentially useful in this respect because they carry out at least two functions. Firstly, they operate at the *textual* level to provide cohesion between the ideas of the text, to indicate conjunctive and/or additive, adversarial, causal and temporal relationships in the text (Schiffrin et al., 2001:55), their function being to "organise propositional information in ways that a perceived audience is likely to find coherent and convincing" (Hyland, 2004:112). The second function is *interpersonal*, whereby they indicate the user's attitude either to the subject matter of the text or to the text itself. In this way metadiscourse markers are textual features at a micro level which provide a significant link to the macro level of a text, reflecting as they do the writer's or speaker's purposes and intentions.

Although aspects of the distinction between the textual and the interpersonal have been challenged (e.g. Hyland & Tse, 2004; Hyland, 2005), the distinction is retained in this project because although all metadiscourse is arguably interactional, in that it is utilised to facilitate the text purpose and the communicative intent of the author, this is less obviously true in the case of unskilled or lower level L2 writers. Bereiter and Scardamalia's (1987) models suggest that unskilled writers in their first language tend to produce text in an additive and relatively unplanned way without extensive consideration of the demands of the task or the expectations of the reader.

Metadiscourse markers in texts are therefore viewed as follows:

- they represent "the language used to express the author's awareness and management of the discourse-as-process" (Burneikaite, 2008:39),
- they do not have an overtly *ideational* function, although they assist in the conveyance of the message,

- they are integral to achieving the outcome or impact that the author intends, through their *interpersonal* and *textual* roles in a text.

2.2. Categories and types of metadiscourse marker

As its starting point, our scheme drew on that of Hyland's (2004) categorisation of metadiscourse items, as the most comprehensive and detailed available. In that scheme, *textual* markers refer to language used to organise the text while *interpersonal* markers manage the social dimensions of the task and allow for commentary on the intended message by the writer. Hyland's initial lists contained over 300 metadiscourse markers, subdivided into ten main categories, with a further four types sub-categorised under *frame markers*. Our categories, adapted from Hyland's, are set out in Table 1.

Table 1: Categories of metadiscourse markers (Adapted from Hyland, 2004:111)

Textual Metadiscourse	**Examples**
Logical connectives	also, although, in addition, on the other hand
Frame markers: Sequencing	finally, firstly, last, to start with
Label stages	all in all, in conclusion, overall, to sum up
Announce goals	I would like to, I will focus on, the aim
Topic shifts	In regard to, now, to come back to, well
Code glosses	for example, in other words, such as, that is to say
Endophoric markers	discussed above, example, section
Evidentials	according to, argue, claim, show
Interpersonal Metadiscourse	
Attitude markers	even, have to, hopefully, important
Hedges	about, could, possible, would
Relational markers	consider, find, let's, you
Person markers	I, me, mine, my
Emphatics	actually, certainly, in fact, must

In our framework, the subcategories of *frame markers* (i.e. *sequencing, label stages, announce goals and topic shifts)* were treated as separate for the purpose of analysis, giving a total of 13 categories.

In addition, items identified as belonging to more than one category or as occurring more frequently in a different category from the one Hyland allocated them to, were adjusted. For example, while 'or' was identified in Hyland's scheme as being a 'code gloss', Waller's (2015) work with a corpus of native-speaker and international student samples found that this word was more commonly used as a logical connector with an additive function (e.g. *humanity does not have to face wars or the problem of starvation*), so it was moved to the category of Logical Connectives.

2.3. Caveats

An important caveat to bear in mind is that metadiscourse markers can offer no more than an initial indication of writer intention and communicative purpose. A second caveat is that even a cursory examination of schemes such as Hyland's (2004) reveals an overlap between categories, with one marker at times having more than one function as exemplified above. Categories can never be entirely watertight (Hyland & Milton, 1997:205). It was therefore an important aspect of our research procedure (see below) that analytical ratings were carefully checked with reference to the discourse context.

2.4. Metadiscourse markers in L2 writing

We now review relevant previous research on metadiscourse to derive our Research Questions (RQ).

a) Quantity of overall markers and interpersonal markers used at different levels

Burneikaite (2008:38) drew on a number of studies examining the use of metadiscourse markers in written texts (e.g. Burneikaite, 2008; Hyland, 2004; Hyland & Tse, 2004; Mauranen, 1993; Vergaro, 2005) and reported little overall difference in the quantity of metadiscourse markers used between L1 and L2 MA theses writers. By contrast, Ädel's (2006) large-scale study comparing L1 and L2 academic writing at university level noted that L2 writers tended to overuse metadiscourse markers, implying weakness in their communicative competence. However, Sanford's (2012) research with a mixed group of relatively novice participants suggested the opposite; as writers develop, they use more such markers, and it is possible that the same is true also for L2 learners. Noble (2010) carried out a small study which explored only textual markers and found that there was an increase in the use of these by more proficient writers at the mid B2 level. In short, the picture for the key question concerning the quantity of metadiscourse markers used by general L2 writers is unclear, so will be explored as a key part of our investigation (see RQ1.1 below).

Furthermore, a small-scale study by Lee and Deakin (2016) indicated that stronger L2 writers used great amounts of *interpersonal* markers in their texts. The study also found that the use of these markers by the stronger L2 writers showed no statistical difference from the native speakers in the corpus. Weaker writers however were found to use fewer *interpersonal markers.* However, the corpus was relatively small with only 25 scripts comprising each group. Burneikaite (2008) and Hyland and Tse (2004) have also suggested that there might also be a difference in the quantity of *interpersonal markers* used by higher and lower level writers.

In light of the contradictions in the research set out above over the amount of metadiscourse used by L2 learners and the small-scale studies which have found a difference in the use of *interpersonal* markers, the first research question was focussed on the quantity of markers overall and the use of *interpersonal markers* (RQ1.2).

b) Variety of markers used at different levels

Intaraprawat and Steffensen (1995:253, emphasis added) looked at a timed essay task that approximated to CEFR B1 and B2 and found that "the good essays showed a greater <u>variety of metadiscourse features</u> within each category than the poor essays". Although their study was small in scale, it suggests a possible pattern of use. Noble (2010) also found a wider range of textual metadiscourse markers in the writing of the stronger students in her small-scale corpus.

In terms of a specific class of markers, namely connectives, Hawkey and Barker (2004) (and also Kennedy et al., 2001) noted the tendency for candidates to overuse these markers. Carlsen's (2010:203) study of metadiscourse markers in the writing of Norwegian learners identified a similar overreliance at lower levels on certain highly-frequent text connectives. She termed these items "connective teddy bears" based on Hasselgren (1994) because they appear to represent an element of security to writers, especially under exam conditions. By contrast, Hawkey and Barker report that higher level writers use fewer of the more common, high frequency connectives, and instead choose a greater variety, including lower-frequency connectives, to carry out the same function.

Since the studies discussed above suggest that a feature of more competent writers is the use of a greater variety of metadiscourse markers, the second research question (RQ2) in this study will examine the differential use of specific classes of markers at different levels, including connectives.

c) Quantity of specific types of markers used at different levels

Burneikaite (2008) noted significant variations in the uses of different categories of markers. While exploration at the level of functional categories can provide some insights into the development of writing at different levels, the differences in the specific linguistic exponents of the functions that writers select at different levels of proficiency may be useful for those developing assessment scales or seeking to develop the writing of learners at these levels. The function of metadiscourse is to signal organisation or stance and therefore the choice of linguistic exponents reflects a writer's perception of text purpose, genre and audience (Badger & White, 2000). For example, the CEFR predicts that at level A2, a writer will be able to link words using *'and', 'but'* and *'because'* but by the end of the B2 level there is a stronger emphasis on discourse skills and it is anticipated that a writer will be able to "link sentence together smoothly into clear, connected discourse using a variety of linking words" (Council of Europe, 2001:29). Precisely which linguistic exponents will be used to do this are not stated in the CEFR, but there is a clear implication that more complex, text-appropriate linkers will be used as learners become more attuned to the notion of audience and text purpose. As stated above, Carlsen's study suggests that learners abandon some of the more high-frequency markers in favour of others which we would anticipate are more complex and more appropriate for the genre. For this reason, RQ3 focuses on changes in the way specific metadiscourse items are used within categories.

Following previous research findings on the overuse or underuse of *Emphatics, Endophorics, Logical connectives* and *Frame markers* in L2 texts (e.g. Burneikaite, 2008; Carlsen, 2010; Hawkey & Barker, 2004), our analysis here will focus only on these categories.

3. Research questions

Drawing on the body of literature reviewed above, this research investigated the following research questions.

RQ 1: *Quantity of markers in general*

RQ1.1 Is there a difference in the overall quantity of metadiscourse markers used by lower and higher level writers?

RQ1.2 Is there a difference in the quantity of interpersonal markers used by higher and lower level writers?

RQ 2: *Variety of markers*

RQ2. Is there a difference in the variety of metadiscourse markers used by higher and lower level writers?

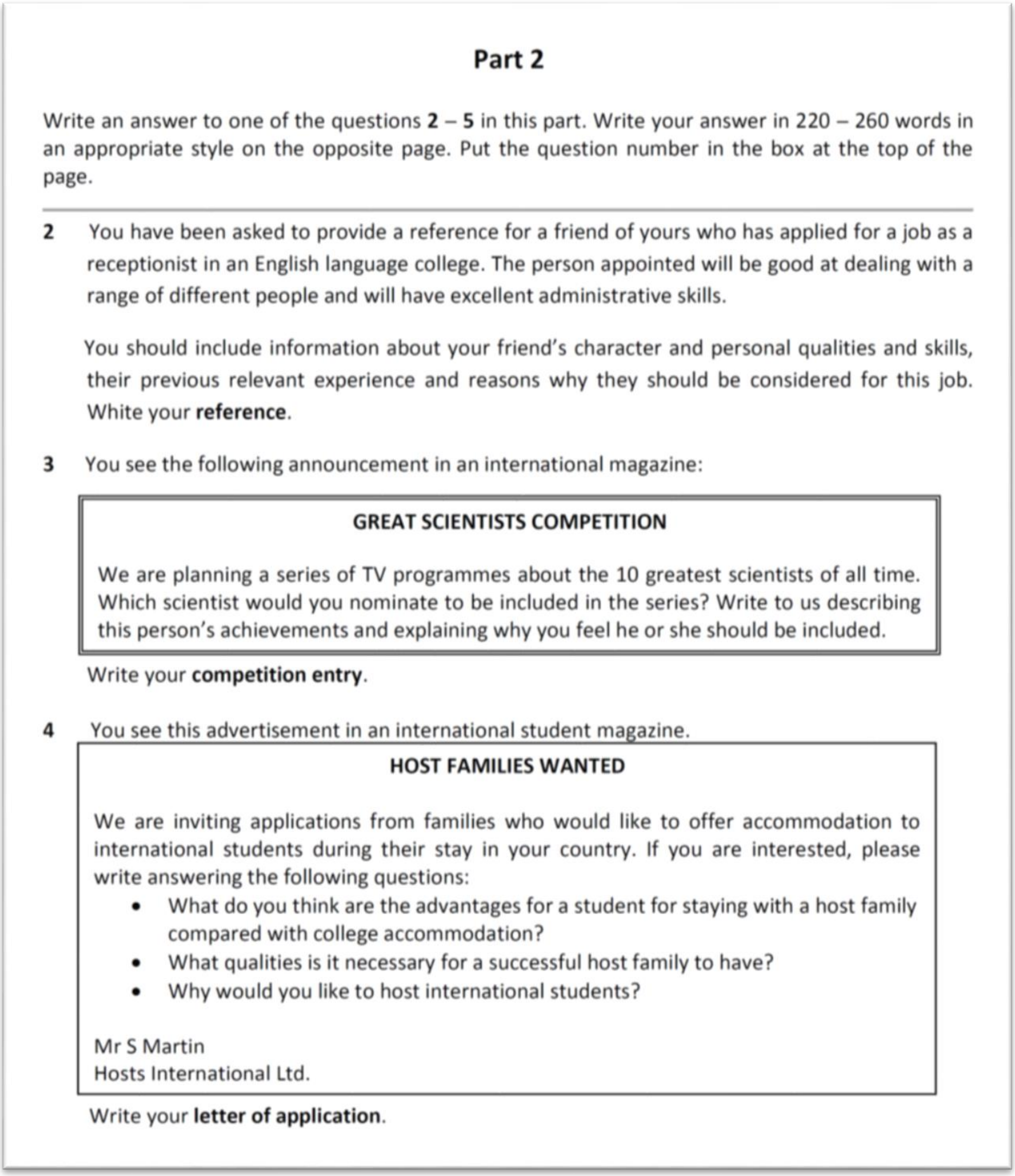
RQ 3: *Individual markers*

RQ3. Is there a difference in the quantities of individual markers in the following categories used by higher and lower level writers: Emphatics, Endophorics, Logical connectives and Frame markers (i.e. Announce goals, Label stages, Sequencing and Topic shifts)?

4. Methodology

4.1. Materials

The initial dataset for the study was a large sample of learner scripts on the three highest levels of the Cambridge English General English examinations, consisting of 1200 pass level scripts, 400 at each of *B2 First, C1 Advanced* and *C2 Proficiency* (formerly called FCE, CAE and CPE). Part 2 of each test, which shares the same response format, was selected for this research. Across the three levels, Part 2 is a situationally-based, knowledge-telling writing task whose situation is specified in no more than 80 words (Shaw & Weir, 2007). In this task, test-takers are given a few choices of question prompts from which they can select one to answer. The prompts were developed to generate different genres of writing such as article, essay, competition entry, letter, report, short story, proposal and review (UCLES, 2012a; 2012b; 2012c; see examples in Figure 1). However, to control the types of writing text to analyse in this study, only the scripts of descriptive, expository and argumentative nature were selected from the pool of 1200 scripts, removing scripts that responded to prompts that required other discourse features (e.g. narrative responses for 'short story' questions). Due to space limitations, only one *C1 Advanced* task is exemplified in Figure 1 (see Shaw & Weir (2007) for detailed task specifications):



Part 2

Write an answer to one of the questions **2 – 5** in this part. Write your answer in 220 – 260 words in an appropriate style on the opposite page. Put the question number in the box at the top of the page.

2 You have been asked to provide a reference for a friend of yours who has applied for a job as a receptionist in an English language college. The person appointed will be good at dealing with a range of different people and will have excellent administrative skills.

You should include information about your friend's character and personal qualities and skills, their previous relevant experience and reasons why they should be considered for this job. White your **reference**.

3 You see the following announcement in an international magazine:

> **GREAT SCIENTISTS COMPETITION**
>
> We are planning a series of TV programmes about the 10 greatest scientists of all time. Which scientist would you nominate to be included in the series? Write to us describing this person's achievements and explaining why you feel he or she should be included.

Write your **competition entry**.

4 You see this advertisement in an international student magazine.

> **HOST FAMILIES WANTED**
>
> We are inviting applications from families who would like to offer accommodation to international students during their stay in your country. If you are interested, please write answering the following questions:
>
> - What do you think are the advantages for a student for staying with a host family compared with college accommodation?
> - What qualities is it necessary for a successful host family to have?
> - Why would you like to host international students?
>
> Mr S Martin
> Hosts International Ltd.

Write your **letter of application**.

Note: Question 5 was omitted in this example due to its required rhetorical mode.

Figure 1: Example *C1 Advanced* Part 2 writing task (UCLES, 2012b:23)

As such, since the rhetorical mode was controlled in the selected dataset, the task difficulty across the three levels differed mainly in relation to topics and text length. All topics are either in social or work domains, but conceptually more complex and lexically more challenging topics are given at higher levels: e.g. ‘friendship’, ‘eating habits’ (*B2 First*), ‘generations living together’, ‘present giving’ (*C1 Advanced*), ‘sportsmen’s pay’, ‘technology’ (*C2 Proficiency*). The required text length is 120-180 words in *B2 First*, 220-260 words in *C1 Advanced* and 280-320 words in *C2 Proficiency*. These tasks are assessed on four criteria: *Content*, *Communicative Achievement*, *Organisation* and *Language*.

Given that a writer’s first language (L1) can play a part in choice of metadiscourse markers (Burneikaite, 2008:45), care was also taken to ensure a balanced mix of L1s in each of the three levels. After detailed preliminary analysis on rhetorical modes and L1 distributions, the dataset was refined to give 900 scripts in total, i.e. 300 scripts at each level. However, despite these efforts to select comparable scripts across the three levels, it should be noted that task differences in terms of topics and text length could potentially be a confounding variable in our study.

4.2. Data Analysis

Most studies of metadiscourse have focussed on clusters or categories of lexical items, e.g. *Hedges*. However, operating purely at the level of categories is potentially misleading, since individual items within a category could distort the figures for other items. Our analysis therefore examined at the macro-level the clusters of items in terms of classes, and also looked at the micro-level at the detailed behaviour of individual elements within each category.

Ädel and Mauranen (2010:9) discuss two traditions in metadiscourse research: the *interactive* and *reflexive* models. The former "uses a broad definition and conceives of metadiscourse as interaction between writer and reader, while [the latter] uses a focused definition and conceives of metadiscourse as a reflexive or metalinguistic function of language". According to this definition, our approach started from the *interactive* tradition, drawing on Hyland's (2004) categorisation of metadiscourse items to seek to analyse large quantities of data. However, to avoid the pitfall in this tradition, namely a tendency to neglect the context in which the metadiscourse markers occur (Ädel & Mauranen, 2010:2−3), our analytical approach drew also on the *reflective* tradition, by ensuring that the contexts in which metadiscourse markers occurred were carefully considered as far as possible. To this end, the analysis was performed in the following six stages.

Stage 1: Analysis tool development

The bespoke computer tool *Text Inspector* (Bax, 2012) was extended to identify over 300 exemplars of metadiscourse markers as classified by Hyland's (2005) and to sort these markers according to the list of 13 categories set out in Table 1. Like other automated text analysis tools, *Text Inspector*'s outputs are not perfectly accurate (e.g. the noun 'will' being misclassified as a verb and wrongly classed as *Emphatic*). To overcome this problem, as well as to avoid the risk of missing other markers, the *Text Inspector* software was designed to allow the analyst to check each example of coding in the context in which it appeared and to alter it or exclude it from the analysis, if necessary.

Stage 2: Automated analysis

All 900 scripts were analysed by *Text Inspector* and outputs were recorded. A cautionary note should be made here that as in many other automated text analysis studies, the software counted all instances of discourse markers regardless of their appropriacies. While whether an item is used context-appropriately is an important consideration in researching L2 learners' use of metadiscourse markers (Bax, Nakatsuhara & Waller, 2013), such studies are usually smaller in scale. The purpose of this large-scale quantitative study was to explore learners' overall use of metadiscourse markers without differentiating appropriate and inappropriate usage.

Stage 3: Manual check to detect computer errors

Three researchers participated in a one-day training session to identify accurately individual metadiscourse markers under the 13 categories. After the definition of each category was explained, six scripts were coded altogether while discussing each occurrence of metadiscourse markers. Another set of six scripts were then coded individually. The individual coding results were compared, and all discrepancies were discussed until a full agreement was reached. Three researchers then independently reviewed the computer outputs for a total of 300 of the texts analysed, 100 for each level, and they used the alteration functionality of *Text Inspector* to propose changes. They then met again to compare the errors that they have identified and agree on their proposed amendments. At this stage, the initial set of over 300 markers was refined to 281. Table 2 sets out the number of lexical items that were agreed to be included in the analysis.

Table 2: Number of lexical items analysed under each metadiscourse category

	Metadiscourse category	**No. of items included**
1	Announce goals (AG)	18
2	Attitude markers (AM)	25
3	Code glosses (CG)	16
4	Emphatics (EM)	38
5	Endophorics (EN)	13
6	Evidentials (EV)	24
7	Hedges (H)	47
8	Label stages (LC)	12
9	Logical connectives (LC)	39
10	Person markers (PM)	4
11	Relational markers (EM)	21
12	Sequencing (S)	16
13	Topic shifts (TS)	8
	Total	**281**

Stage 4: Adjustments to the automated analysis

On the basis of these amendments, the *Text Inspector* outputs for the whole set of texts was manually adjusted to mitigate any biases in the computer analysis. Although the amendments of the remaining 600 texts were based on estimates from the analysis of the 300 texts, it was hoped that a more accurate picture of the incidence of metadiscourse markers was obtained by these amendments. This human checking and adjustments formed an important part of the research project, seeking to overcome the problem noted above of a failure to consider the discourse context, as recommended by Hyland and Milton (1997:205).

Stage 5: Data standardisation

As mentioned earlier, candidates in this project produced different lengths of expository essays at *B2 First, C1 Advanced* and *C2 Proficiency*. Since comparing the raw datasets based solely on word count would confound the results, the raw data was standardised by computing 'frequency per 100 words' so that the results would not be distorted by the length of scripts collected from the three exams. All statistical analyses to be explained below were performed on the ratio dataset standardised in this way. Table 3 and Figure 2 show that although the raw count of metadiscourse markers increased as the level went up, when the figures were standardised, the number of metadiscourse markers per 100 words decreased as the level went up. It should be noted that although we follow the assumption that the decreasing trend in the transformed dataset is essentially the function of candidates' proficiency level, we cannot eliminate the possibility that even the same writer might vary in the frequency of metadiscourse marker usage depending on the length of the texts. This is one of the limitations of the study.

Table 3: Mean frequency of lexical tokens and metadiscourse markers

	B2 First **(*N*=300)**	***C1 Advanced*** **(*N*=300)**	***C2 Proficiency*** **(*N*=300)**
Tokens per script	192.69	279.92	365.98
Metadiscourse markers per script	34.05	44.61	46.79
Metadiscourse markers per 100 words	17.67	15.93	12.74

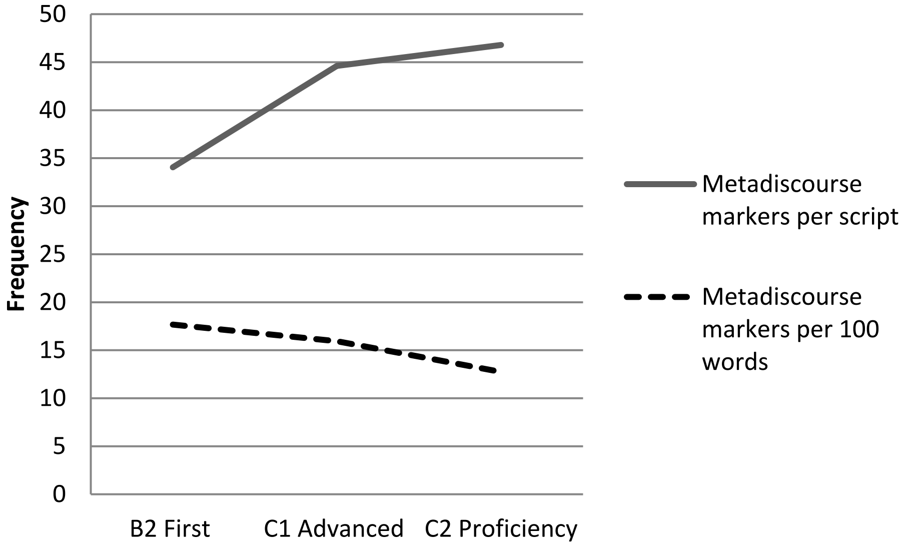


Figure 2: Use of metadiscourse markers across the three levels

Stage 6: Statistical analyses

Using the standardised dataset, the use of metadiscourse markers in the 900 scripts was statistically compared across the three levels to address the three research questions of this study. Given non-normal distributions of the dataset, non-parametric tests were used for inferential statistics. Kruskal-Wallis tests followed by Mann-Whitney *U* tests were run to compare the overall use of metadiscouse markers (RQ1) and to examine the range of markers used (RQ2). Employing the same statistical methods, the use of individual items within selected metadiscourse classes was analysed (RQ3). The Kruskal-Wallis tests were used to identify overall differences across the three levels and Mann-Whitney *U* tests served to identify differences in pairwise post-hoc comparisons (Field, 2009:565). In order to avoid Type I errors, the Bonferroni adjustment was applied to Mann-Whitney *U* results.

5. Results and discussion

5.1. Quantity of metadiscourse markers used across levels (RQ1)

Table 4 presents the results of the Kruskal-Wallis tests which examined whether there was any overall difference in the use of metadiscourse markers in the 13 categories collectively and individually across the three levels.

Table 4: Kruskal-Wallis tests on aggregated numbers of metadiscourse makers in each category

	Median (per 100 words)			**X^2**	***p***
	B2 First	**C1 Advanced**	**C2 Proficiency**		
Announce goals	0.00	0.00	0.00	1.07	0.59
Attitude markers	0.52	0.36	0.55	5.59	0.06
Code glosses	0.00	0.00	0.03	4.83	0.09
Emphatics	1.76	1.79	1.51	25.87	0.00
Endophorics	0.00	0.00	0.00	20.87	0.00
Evidentials	0.00	0.00	0.00	48.65	0.00
Hedges	1.56	1.43	1.37	9.71	0.01
Label stages	0.00	0.00	0.00	6.87	0.03
Logical connectives	5.29	5.15	5.28	2.81	0.25
Person markers	1.56	0.71	0.41	71.58	0.00
Relational markers	4.15	4.29	1.37	146.71	0.00
Sequencing	0.00	0.00	0.15	4.37	0.11
Topic shifts	0.00	0.00	0.00	8.00	0.02
Total	**16.66**	**15.47**	**12.29**	**147.68**	**0.00**

Addressing RQ1.1, the last line of Table 4 shows that there was a significant overall difference for the total use of metadiscourse markers across the levels. Descriptive statistics also suggest some types of metadiscourse markers were more extensively used than others. The extensive use of *Logical connectives* and *Relational markers* was particularly noteworthy. Significant overall differences were obtained for eight metadiscourse categories: *Emphatics, Endophorics, Evidentials, Hedges, Label stages, Person markers, Relational markers,* and *Topic shifts*.

To elucidate where these differences originated from, Mann-Whitney *U* tests were then carried out for the eight categories as well as overall use. Table 5 summarises the results of the post-hoc comparisons (see Appendix A for more details), indicating whether the use of each marker remained the same, or increased or decreased at each level.

Table 5: Summary of post-hoc comparisons across three proficiency levels

	Type*	***B2 First*** Median	dif. **	***C1 Advanced*** Median	dif. **	***C2 Proficiency*** Median	dif. **	***B2 First*** Median	Pattern of change
AG (n.s.)	T	0.00	=	0.00	=	0.00	=	0.00	A
AM (n.s.)	I	0.52	=	0.36	=	0.55	=	0.52	A
CG (n.s.)	T	0.00	=	0.00	=	0.03	=	0.00	A
EM	I	1.76	=	1.79	>	1.51	<	1.76	C
EN	T	0.00	=	0.00	<	0.00	>	0.00	B
EV	T	0.00	<	0.00	≤	0.00	>	0.00	B
H	I	1.56	=	1.43	>	1.37	=	1.56	C
LS	T	0.00	>	0.00	=	0.00	=	0.00	C
LC (n.s.)	T	5.29	=	5.15	=	5.28	=	5.29	A
PM	I	1.56	>	0.71	>	0.41	<	1.56	C
RM	I	4.15	=	4.29	>	1.37	<	4.15	C
S (n.s.)	T	0.00	=	0.00	=	0.15	=	0.00	A
TS	T	0.00	=	0.00	=	0.00	<	0.00	C
Total		**16.66**	**>**	**15.47**	**>**	**12.29**	**<**	**16.66**	**C**

* Note. T: textual, I: interpersonal
** Note. =: No sig difference; < >: Sig difference at 0.017 with a Bonferroni adjustment;
≤ ≥: Sig difference at 0.05

The last line of Table 5 suggests that more proficient L2 writers made significantly less use of metadiscourse markers than writers at lower levels. One of the possible reasons for this finding could be that high level L2 writers in our sample might be deploying other, non-explicit discourse strategies instead in order to achieve their ends (Waller, 2015). As indicated in the final column of Table 5, we can broadly classify the 13 metadiscourse categories as well as overall use into the following three patterns in terms of usage by candidates at the three levels of proficiency.

- **A.** No significant difference across the three levels: *Announce goals, Attitude markers, Code glosses, Logical connectives, Sequencing*
- **B.** Increased use as the level goes up: *Endophorics, Evidentials*

C. Decreased use as the level goes up: *Emphatics, Hedges, Label stages, Person markers, Relational markers, Topic shifts* and overall use of markers as a whole

The second part of RQ1 concerned the difference in the quantity of *interpersonal markers* used by higher and lower level writers (RQ1.2). To address this, we examined the quantity of interpersonal markers as a whole against that of textual markers across the three proficiency levels. The second column of Table 5 shows which of the 13 metadiscourse categories are classified into the two types of markers. The Kruskal-Wallis tests indicated that there was a significant difference for the use of interpersonal markers across the three levels ($X^2(2)=188.42$, $p<0.001$), but not for textual markers ($X^2(2)=4.36$, $p=0.113$). Subsequently, Mann-Whitney *U* tests indicated that the use of interpersonal markers differed between all levels at the significant level of 0.017, as shown in Table 6.

Table 6: Post-hoc comparisons for Interpersonal markers

B2 First	>	C1 Advanced	>	C2 Proficiency	<	B2 First
Median= 10.90	U=38100.50 W=83250.50 Z=-3.25 p=.001	**Median= 9.65**	U=22004.00 W=67154.00 Z=-10.83 p<.001	**Median= 5.96**	U=18529.00 W=63679.00 Z=-12.47 p<.001	**Median= 10.90**

Note. < >: Sig difference at 0.017

Therefore, unlike previous research (Burneikaite, 2008; Hyland & Tse, 2004), the present data showed that significantly fewer interpersonal markers were used as proficiency levels increased, while the use of textual markers did not show any particular patterns across the three levels. This is an interesting finding since it suggests that higher level writers in this genre choose a more impersonal style, perhaps in line with their perception of the requirements of the academic discourse community of which they are aiming to be a part.

5.2. Variety of markers used across levels (RQ2)

We then examined how many different unique metadiscourse markers in each of the 13 metadiscourse categories were used by each individual at least once. As shown in Table 7, there was a significant overall difference for 10 metadiscourse categories: *Attitude markers, Code glosses*, *Emphatics*, *Endophorics*, *Evidentials*, *Hedges*, *Logical connectives, Person markers, Relational markers*, and *Topic shifts*, while the remaining three showed no significant differences. Mann-Whiteney *U* tests revealed that in 8 of the 10 categories (*Attitude markers, Code glosses*, *Emphatics*, *Endophorics*, *Evidentials*, *Hedges*, *Logical connectives*, *Topic shifts*) showed that higher level writers used a greater variety of metadiscourse markers than lower level writers, increasing from *B2 First* to *C2 Proficiency* significantly at each stage; *Relational markers* increased at *C1 Advanced* then decreased at *C2 Proficiency*; *Person markers* showed a significant decrease across levels (see ‘Summary of Mann-Whitney *U* test’ in Table 7 and Appendix B).

Table 7: The number of unique metadiscourse markers used by each individual

Metadiscourse category	Level	Median	Mean	SD	Kruskal-Wallis test		Summary of Mann-Whitney *U* test*
					X^2	*p*	
Announce goals	B2	0.00	0.06	0.23	.907	.635	-
	C1	0.00	0.04	0.20			
	C2	0.00	0.05	0.24			
Attitude Markers	B2	1.00	1.08	0.87	21.333	.000	B2 < C1 C1 ≤ C2 B2 < C2
	C1	1.00	1.29	0.96			
	C2	1.00	1.46	0.96			
Code Glosses	B2	0.00	0.53	0.67	62.798	.000	B2 < C1 C1 = C2 B2 < C2
	C1	1.00	0.98	0.83			
	C2	1.00	1.10	1.06			
Emphatics	B2	3.00	3.16	1.62	40.790	.000	B2 < C1 C1 = C2 B2 < C2
	C1	4.00	3.86	1.81			
	C2	4.00	4.09	1.90			
Endophorics	B2	0.00	0.03	0.18	23.081	.000	B2 = C1 C1 < C2 B2 < C2
	C1	0.00	0.06	0.24			
	C2	0.00	0.14	0.38			
Evidentials	B2	0.00	0.18	0.46	59.280	.000	B2 < C1 C1 ≤ C2 B2 < C2
	C1	0.00	0.45	0.69			
	C2	0.00	0.59	0.79			
Hedges	B2	2.00	2.39	1.50	97.259	.000	B2 < C1 C1 ≤ C2 B2 < C2
	C1	3.00	3.35	1.60			
	C2	4.00	3.69	1.76			
Label stages	B2	0.00	0.26	0.45	4.153	.125	-
	C1	0.00	0.21	0.41			
	C2	0.00	0.28	0.46			
Logical connectives	B2	5.00	5.00	1.65	129.297	.000	B2 < C1 C1 < C2 B2 < C2
	C1	6.00	6.03	1.63			
	C2	7.00	6.93	2.29			
Person markers	B2	1.00	1.51	1.02	27.409	.000	B2 ≥ C1 C1 > C2 B2 > C2
	C1	1.00	1.34	1.04			
	C2	1.00	1.08	1.01			
Relational markers	B2	2.00	2.54	1.31	82.958	.000	B2 < C1 C1 > C2 B2 = C2
	C1	4.00	3.64	1.60			
	C2	3.00	2.73	1.67			
Sequencing	B2	0.00	0.75	0.95	7.568	.023	-
	C1	0.00	0.72	0.98			
	C2	1.00	0.91	1.06			
Topic shifts	B2	0.00	0.25	0.49	36.752	.000	B2 < C1 C1 < C2 B2 < C2
	C1	0.00	0.40	0.55			
	C2	0.00	0.53	0.63			

*Note. =: No sig difference; < >: Sig difference at 0.017; ≤ ≥: Sig difference at 0.05

5.3. Analysis individual items in specific classes of markers (RQ3)

We now address RQ3 by analysing individual items within these categories, going beyond most previous studies which tend only to consider the categories as a whole. The analysis focused only on *Emphatics, Endophorics, Logical connectives* and four *Frame markers.*

Table 8 presents the metadiscourse markers which were selected for this analysis, identified as good representatives of each category; those excluded were used only rarely in our data. It also sets out the tendency of each category (e.g. *Announce goals* showed no significant difference between *B2 First, C1 Advanced* and *C2 Proficiency*, whereas the category of *Emphatics* showed a significant decrease in use). The final column gives details of the behaviours of particular markers of interest (see Appendix C for the results of Kruskal-Wallis tests and summaries of post-hoc tests).

Table 8: Selected items for individual lexical analysis

Metadiscourse category	No. of total items	No. of selected items	Selected items for analysis	Tendency as a category, as levels increase	Notes
Announce goals (frame markers)	18	1	*I would like to*	No significant difference	The only item selected in the category showed no significant difference across the three levels, probably due to the short length of the texts, and the fact that the goals were already to a large extent set for candidates under exam conditions.
Emphatics	38	21	*actually, always, certainly, clearly, definitely, essential, even if, I believe, in fact, indeed, know, must, never, obviously, of course, should, sure, the fact that, undoubtedly, will, won't*	Decrease	A few items with large differences skewed the picture downwards for the category as a whole; in the case of *always* there was a large decrease in use from C1 to C2, and in the cases of *know* and *will*, there was a significant increase from B2 to C1, then a significant decrease at C2. The use of 7 of the 21 items increased as the level went up, although the actual differences were very limited except in the case of *should*.

Endophorics	13	1	*example*	Increase	Given that the texts here were very short, it was anticipated that there would not be much use of endophoric markers, and this proved to be the case. Only one (*example*) was used sufficiently to be explored further, and showed an increase as the level increased.
Label stages (frame markers)	12	5	*all in all, in conclusion, overall, to conclude, to sum up*	Decrease	Use of this class was small, probably because the texts were relatively short, with correspondingly few major sections which would call for labelling.
Logical connectives	39	24	*also, although, and, as a result, because, besides, but, consequently, even though, furthermore, however, in addition, moreover, nevertheless, on the contrary, on the other hand, since, so, therefore, though, thus, whereas, while, yet*	No significant difference	14 of these items showed significant increases as the level went up, many items showing median values of 0, meaning that most scripts did not use that item at all. By contrast, the decreasing tendency or a decreasing-increasing tendency shown by *because*, *but* and *and* were more marked.
Sequencing (frame markers)	16	9	*finally, first, firstly, last, lastly, next, secondly, thirdly, to start with*	No significant difference	Items showed no significant changes across levels, with the small exception of a significant difference in the use of *first* from C1 to C2. On the whole, this marker is not used differentially across the three levels perhaps because relatively short texts there do not need major sequencing moves.
Topic shifts (frame markers)	8	2	*now, well*	Decrease	The first of these (*now*) showed no difference in use when used across levels; the second (*well*) showed a decrease across levels, but the data are not sufficient to draw useful conclusions about the class as a whole.

Table 8 allows us to address RQ3. Starting with *Emphatics*, there was a clear difference in the quantity used by higher and lower level writers, but the direction of the difference was rather mixed. While the use of some emphatics (e.g. *always*) declined as the level went up, the use of many other emphatics increased with higher level writers, in line with Burneikaite's

(2008) findings. The rather mixed picture confirms the importance of analysing trends within each category in a more nuanced way, rather than simply accepting the overall figure for the category as a whole. In terms of learner behaviour, it suggests that in general writers do use certain emphatics more frequently, often apparently more subtle or complex ones, as their level increases, and at *C2 Proficiency* they also reject those which they consider to be more straightforward or basic. This is in line with the trend towards more sophisticated markers in the classes already discussed.

Regarding *Endophoric* markers, our data did not allow a clear answer to the question regarding the quantity of these used by higher and lower level writers, perhaps owing to the brevity of the texts involved, except to say that there was no evidence that our higher level writers do use more endophorics in the way which Burneikaite (2008) implied. This may also be a result of the task type; candidates in these types of examination task (i.e. knowledge-telling task) are required to write from their own resources and do not have texts to refer to.

Regarding *Logical connectives*, Hawkey and Barker's (2004) and Carlsen's (2010) suggestion that higher level writers show significantly lower use of more common logical connectives, a key element of CEFR higher levels according to Carlsen, seems to be clearly supported by our data. Participants in our project clearly used fewer of the very high frequency (K1 of the British National Corpus) logical connectives which are commonly highlighted by text books and teachers as their levels increased, and started to use more elaborate forms more frequently.

It is worth noting that scripts at *B2 First* and *C1 Advanced* levels were often indistinguishable in the use of this type of marker. In 12 of the 17 items with significant differences, *B2 First* and *C1 Advanced* were not distinguishable; the significant difference in those cases emerging only when *C2 Proficiency* is taken into account (see Appendix C and Bax et al. (2013) for details of post-hoc results). This may indicate that there is a bigger step change in the use of connectives between levels C1 and C2 than before C1.

It is also noteworthy that although this class showed no significant movement as a whole, in the case of two of the markers (*but¸ because*), which are arguably conceptually relatively simple, there was a clear decrease in use across levels. By contrast, markers in the larger group which demonstrated a clearly increased use (14 in number) are arguably more subtle and complex conceptually. It is interesting to note that this appears congruent with the evidence from the English Vocabulary Profile (EVP) wordlists developed on the basis of large quantities of learner scripts (Cambridge EVP, nd), in which many of these markers are seen to be used only at higher CEFR levels.

With reference to *Frame markers*, the picture was rather mixed. In terms of the four sub-categories, two showed no significant difference across levels (*Announce goals, Sequencing*), whereas the other two did not provide conclusive data (*Label stages and Topic shifts*). Perhaps owing to the brevity of the texts used in the study, it was not possible to agree with Burneikaite's (2008) suggestion that higher level students use decreasing quantities of *frame markers*, at least in exam writing of the kind studied here.

6. Conclusions

The findings of this study have provided a number of insights into the use of metadiscourse markers by writers at the levels of the CEFR investigated in this study. Firstly, more proficient L2 writers in this study used significantly *fewer* number of metadiscourse markers than writers at lower levels (RQ1.1). This is congruent with Ädel's (2006) study which compared L1 and L2 academic writing, while challenging Sanford's (2012) research with novice L2 writers. Given the relatively high level learners targeted in this study (i.e. B2-C2), it could be interpreted that the overall use of metadiscourse markers increases as learners acquire (mainly basic) metadiscourse markers, but after reaching a certain level, the use of explicit metadiscourse markers decreases as they learn more sophisticated and subtle ways to express the organisation of a text without heavily depending on explicit markers.

In terms of the *interpersonal* versus *textual* use of markers (RQ1.2), the data from this study runs contrary to Burneikaite (2008) and Hyland and Tse (2004), with significantly fewer interpersonal markers used at higher levels of proficiency. Textual makers by contrast did not show any particular patterns across the three levels. The finding on interpersonal markers is of interest as it might be indicative of learners demonstrating more concern regarding meeting the expectations of genre and audience the adoption of an impersonal style; a development predicted by the CEFR and its descriptors.

Furthermore, there were changes in the categories of metadiscourse markers used (RQ2) with writers at higher levels overall displaying a greater range of markers to fulfil the different metadiscourse functions. This is consistent with the findings of Intaraprawat and Steffensen (1995) as well as verifying the CEFR's prediction that the range of discourse markers will be higher in the writing of more proficient learners.

Regarding the actual linguistic exponents used for different functions (RQ3), the study found some evidence of writers at higher levels abandoning the highest-frequency 'lexical teddy bears' in favour of other makers. This would also be consistent with Bereiter and Scardamalia's (1987) models, in that it suggests that higher level writers are better able to plan what they write and therefore do not need to rely so much on simple markers such as *and* or *because* unlike writers at lower levels who have a more additive approach to text production.

Hence, the study confirms many of the suggestions made by the CEFR regarding the development of discourse as a key part of higher levels of language proficiency.

Potential limitations include elements of comparability across levels in terms of text difficulty despite our attempts to control for it. However, a positive consequence of this is that all learners performed tasks that were targeted to their proficiency levels. It should also be noted that a repeated-measure design study is necessary to examine how writers use metadiscourse markers in different lengths of texts, in order to confirm the validity of the way in which our data were standardised.

The current study also only included learners of English and not native speakers in the texts examined. However, the purpose of the study was to examine how metadiscourse markers were used by writers at different levels of proficiency rather than looking at native speaker/non-native speaker variation. The inclusion of learners at C2, a level at which learners can produce "clear, smoothly flowing text in an appropriate style" (Council of Europe, 2001:27) means that there is already a comparison with a group who displays highly proficient and educated writing skills, arguably higher than that of many native speakers. This is also a level of proficiency which learners could realistically aspire to reach.

Despite these limitations, we believe that this large-scale study has provided an important new insight into the ways in which learners develop their use of metadiscourse markers in their writing as they progress up the levels, and thereby into the ways in which their sense of discourse as a whole develops through time. Our study, we would argue, therefore contributed to the debate about how discourse is to be viewed in terms of the CEFR as a whole, and this in turn could impact significantly on the activities of teachers and learners, of exam boards, materials writers and others involved in English language education. Furthermore, a vital part of this research was the demonstration of the importance of analysing not only the behaviour of whole classes of metadiscourse items, but the behaviour or individual members of each class, since our data demonstrated that a single item could significantly skew the pattern of the class as a whole. This was a crucial element of the study and one that has been missing in many previous studies.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English.* Amsterdam/Philadelphia: John Benjamins Publishing

Ädel, A., & Mauranen, A. (2010). Metadiscourse: Diverse and divided perspectives. *Nordic Journal of English Studies, 9*(2), 1-11.

Badger, R., & White, G. (2000). A process genre approach to teaching writing. *ELT Journal, 54*(2),153-160.

Bax, S. (2012). *Text Inspector*. Online text analysis tool. Available at https://textinspector.com/

Bax, S., Nakatsuhara, F., & Waller, D. (2013). *Computer-based analysis of metadiscourse in candidates' writing at Cambridge FCE, CAE, CPE levels.* Project report submitted to the Cambridge English Language Assessment.

Bhatia, V. (1993). *Analysing genre: Language use in professional settings.* London: Longman.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Burneikaite, N. (2008). Metadiscourse in linguistics master's theses in English L1 and L2. *Kalbtyra*, *59*(3), 38-46.

Cambridge EVP (nd.). http://www.englishprofile.org/index.php/resources/wordlists Accessed 30/7/2013.

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment.* Cambridge: Cambridge University Press.

Crismore, A., & Farnsworth, R. (1990). Metadiscourse in popular and professional science discourse. In W. Nash (Ed.) *The Writing Scholar: Studies in academic discourse*. Newbury Park, CA: Sage.

Crompton, P. (2012). Characterising hedging in undergraduate essays by Middle-Eastern students. *The Asian ESP Journal*, 8(2), 55-78.

Field, A. (2009) *Discovering Statistics using SPSS: (And sex and drugs and rock 'n' roll). 3rd ed.* London: SAGE.

Harris, Z. S. (original work published in 1959). Linguistic transformations information retrieval. In Z. Harris (1970). *Papers in structural and transformational linguistics* (pp.253-277). Dordrecht: D. Reidel.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the way Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*, 237-258.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*, 122-159.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing.* Michigan: University of Michigan Press.

Hyland, K. (2005). *Metadiscourse*. London, Continuum.

Hyland, K., & J. Milton. (1997). Hedging in LI and L2 student writing, *Journal of Second Language Writing, 6*(2), 183-206.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics, 25*(2), 156 – 177.

Intarpaprawat, P., & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*(3), 253 – 272.

Kennedy, C, T., Dudley-Evans T., & Thorp, D. (2001). *Investigation of linguistic output of writing task two*. British Council.

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.

Mauranen, A. (1993). *Cultural differences in academic rhetoric: A textlinguistic study*. Frankfurt am Main: Peter Lang.

McCarthy, M., & Carter, R. (1994). *Language as discourse: Perspectives for language teaching.* London: Longman.

Noble, W. (2010). Understanding metadiscoursal use: Lessons from a 'local' corpus of learner academic writing. *Nordic Journal of English Studies, 9*(2), 145-169.

Ozdemir, N. O., & Longo, B. (2014). Metadiscourse Use in Thesis Abstracts: A Cross-cultural Study. *Procedia-Social and Behavioural Sciences, 141*, 59-63.

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

Schiffrin, D., Tannen, D., & Hamilton, H. (2001). *The handbook of discourse analysis*. Oxford: Blackwall Publishing.

Skulstad, A. S. (2005). The use of metadiscourse in introductory sections of a new genre, *International Journal of Applied Linguistics, 15,* 71-86.

Shaw, S., & Weir, C. J. (2007). *Examining writing and practice in assessing second language writing. Studies in Language Testing 26.* Cambridge: UCLES/Cambridge University Press.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

UCLES. 2012a. *Cambridge English First (FCE): Handbook for teachers,* downloaded from http://www.cambridgeenglish.org/ [accessed 22/01/13]

UCLES. 2012b. *Cambridge English Advanced (CAE): Handbook for teachers*, downloaded from http://www.cambridgeenglish.org/ [accessed 22/01/13]

UCLES. 2012c. *Cam ridge English Proficiency (CPE): Handbook for teachers*, downloaded from http://www.cambridgeenglish.org/ [accessed 22/01/13]

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse, *College Composition and Communication 36,* 82 – 93.

Vergaro, C. (2005). Dear Sirs, I hope you will find this information useful: Discourse strategies in Italian and English 'For Your Information' (FYI) letters. *Discourse Studies, 7*(1), 109-135.

Waller, D. (2015). *Investigation into the features of written discourse at levels B2 and C1 of the CEFR.* Unpublished PhD dissertation, University of Bedfordshire, UK.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 1 – 20.

Appendix A: Post-hoc comparisons on the number of metadiscourse markers (per 100 words)

	Emphatics	Endophorics	Evidentials	Hedges	Label stages	Person markers	Relational markers	Topic shifts	Total
Comparisons between *B2 First* and *C1 Advanced*									
Mann-Whitney U	41633.000	43882.000	36658.000	41912.500	40327.500	33737.000	43131.500	41842.000	37235.000
Wilcoxon W	86783.000	89032.000	81808.000	87062.500	85477.500	78887.000	88281.500	86992.000	82385.000
Z	-1.588	-1.441	-5.120	-1.457	-2.994	-5.338	-0.880	-1.841	-3.657
p	0.112	0.150	0.000	0.145	0.003	0.000	0.379	0.066	0.000
Comparisons between *C1 Advanced* and C2 *Proficiency*									
Mann-Whitney U	34241.500	41938.000	41168.000	38168.000	44170.000	36494.000	21158.500	43209.000	26050.500
Wilcoxon W	79391.500	87088.000	86318.000	83318.000	89320.000	81644.000	66308.500	88359.000	71200.500
Z	-5.076	-2.814	-2.049	-3.222	-.525	-4.068	-11.235	-.945	-8.925
p	0.000	0.005	0.041	0.001	0.600	0.000	0.000	0.344	0.000
Comparisons between *B2 First* and *C2 Proficiency*									
Mann-Whitney U	38101.500	40642.000	33251.000	41929.000	42822.000	28309.500	24511.500	39311.000	20900.000
Wilcoxon W	83251.500	85792.000	78401.000	87079.000	87972.000	73459.500	69661.500	84461.000	66050.000
Z	-3.252	-4.284	-6.907	-1.449	-1.326	-7.947	-9.656	-3.164	-11.351
p	0.001	0.000	0.000	0.147	0.185	0.000	0.000	0.002	0.000

Appendix B: Post-hoc comparisons on the number of unique metadiscourse markers used by each individual (per 100 words)

	Attitude markers	Code glosses	Emphatics	Endophorics	Evidentials	Hedges	Logical connectives	Person markers	Relational markers	Topic shifts
Comparisons between *B2 First* and *C1 Advanced*										
Mann-Whitney U	39811.000	31225.000	35206.000	43800.000	35543.000	29404.000	29081.000	40694.000	26539.000	38740.000
Wilcoxon W	84961.000	76375.000	80356.000	88950.000	80693.000	74554.000	74231.000	85844.000	71689.000	83890.000
Z	-2.582	-7.031	-4.685	-1.547	-5.847	-7.482	-7.631	-2.106	-8.855	-3.620
p	.010	.000	.000	.122	.000	.000	.000	.035	.000	.000
Comparisons between *C1 Advanced* and *C2 Proficiency*										
Mann-Whitney U	40885.000	43856.500	41714.000	41691.000	41360.000	40111.000	34991.000	38533.000	30845.000	39980.500
Wilcoxon W	86035.000	89006.500	86864.000	86841.000	86510.000	85261.000	80141.000	83383.000	75695.000	84830.500
Z	-2.031	-.570	-1.502	-3.045	-1.977	-2.344	-4.778	-3.107	-6.710	-2.578
p	.042	.569	.133	.002	.048	.019	.000	.002	.000	.010
Comparisons between *B2 First* and *C2 Proficiency*										
Mann-Whitney U	35701.000	31572.000	32154.500	40495.000	32239.000	25784.000	22617.000	34246.500	42472.000	34224.000
Wilcoxon W	80851.000	76722.000	77304.500	85645.000	77389.000	70934.000	67767.000	79096.500	87622.000	79374.000
Z	-4.610	-6.817	-6.081	-4.432	-7.560	-9.194	-10.667	-5.203	-1.146	-6.028
p	.000	.000	.000	.000	.000	.000	.000	.000	.252	.000

Appendix C: Kruskal-Wallis tests and post-hoc summaries on the use of items selected for individual analyses (See Bax et al., 2013 for post-hoc test figures)

Announce goal

Item	Level	Median	Mean	SD	X^2	*p*	Post-hoc summary	Pattern of change			
								no diff	-	-	-
I would like to	B2	0.000	0.028	0.151	0.399	0.819	-	✓			
	C1	0.000	0.013	0.073							
	C2	0.000	0.015	0.073							

Emphatics

Item	Level	Median	Mean	SD	X^2	*p*	Post-hoc summary	Pattern of change			
								no diff	up	down	up-down
clearly	B2	0.000	0.005	0.052	4.730	0.094	-	✓			
	C1	0.000	0.004	0.036							
	C2	0.000	0.008	0.047							
definitely	B2	0.000	0.012	0.089	3.493	0.174	-	✓			
	C1	0.000	0.012	0.076							
	C2	0.000	0.013	0.058							
even if	B2	0.000	0.052	0.167	1.765	0.414	-	✓			
	C1	0.000	0.043	0.148							
	C2	0.000	0.021	0.076							
I believe	B2	0.000	0.035	0.155	6.212	0.045	-	✓			
	C1	0.000	0.013	0.067							
	C2	0.000	0.029	0.103							
in fact	B2	0.000	0.035	0.136	0.391	0.822	-	✓			
	C1	0.000	0.020	0.083							
	C2	0.000	0.015	0.063							
must	B2	0.000	0.123	0.348	3.140	0.208	-	✓			
	C1	0.000	0.089	0.219							
	C2	0.000	0.096	0.223							
never	B2	0.000	0.100	0.234	0.208	0.901	-	✓			
	C1	0.000	0.079	0.191							
	C2	0.000	0.066	0.154							
obviously	B2	0.000	0.007	0.060	5.813	0.055	-	✓			
	C1	0.000	0.015	0.073							
	C2	0.000	0.013	0.058							
of course	B2	0.000	0.067	0.185	3.127	0.209	-	✓			
	C1	0.000	0.070	0.167							
	C2	0.000	0.062	0.131							
sure	B2	0.000	0.059	0.175	0.632	0.729	-	✓			
	C1	0.000	0.045	0.139							
	C2	0.000	0.028	0.086							
won't	B2	0.000	0.029	0.127	5.001	0.082	-	✓			
	C1	0.000	0.020	0.083							
	C2	0.000	0.007	0.049							
actually	B2	0.000	0.024	0.110	11.528	0.003	B2 < C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.046	0.149							
	C2	0.000	0.040	0.107							
certainly	B2	0.000	0.016	0.089	10.564	0.005	B2 < C1		✓		

	C1	0.000	0.029	0.101			C1 = C2				
	C2	0.000	0.029	0.090			B2 < C2				
indeed	B2	0.000	0.012	0.099	10.930	0.004	B2 = C1		✓		
	C1	0.000	0.012	0.064			C1 = C2				
	C2	0.000	0.021	0.079			B2 < C2				
essential	B2	0.000	0.007	0.060	7.835	0.020	B2 = C1		✓		
	C1	0.000	0.010	0.058			C1 = C2				
	C2	0.000	0.015	0.067			B2 < C2				
should	B2	0.000	0.180	0.389	32.799	0.000	B2 < C1		✓		
	C1	0.000	0.276	0.401			C1 = C2				
	C2	0.000	0.261	0.371			B2 < C2				
the fact that	B2	0.000	0.019	0.107	48.691	0.000	B2 < C1		✓		
	C1	0.000	0.044	0.128			C1 < C2				
	C2	0.000	0.082	0.165			B2 < C2				
undoubtedly	B2	0.000	0.002	0.030	6.134	0.047	B2 ≤ C1		✓		
	C1	0.000	0.008	0.054			C1 = C2				
	C2	0.000	0.008	0.047			B2 < C2				
always	B2	0.208	0.204	0.227	40.682	0.000	B2 = C1			✓	
	C1	0.000	0.248	0.361			C1 > C2				
	C2	0.000	0.111	0.205			B2 > C2				
know	B2	0.000	0.061	0.118	25.431	0.000	B2 < C1				✓
	C1	0.000	0.154	0.262			C1 > C2				
	C2	0.000	0.062	0.140			B2 < C2				
will	B2	0.000	0.325	0.525	29.032	0.000	B2 < C1				✓
	C1	0.357	0.411	0.463			C1 > C2				
	C2	0.000	0.256	0.384			B2 = C2				

Endophorics

Item	Level	Median	Mean	SD	X^2	*p*	Post-hoc summary	Pattern of change			
								-	up	-	-
example	B2	0.000	0.014	0.103	23.355	0.000	B2 ≤ C1		✓		
	C1	0.000	0.021	0.090			C1 < C2				
	C2	0.000	0.040	0.120			B2 < C2				

Label stages

Item	Level	Median	Mean	SD	X^2	*p*	Post-hoc summary	Pattern of change			
								no diff	up	down	down-up
overall	B2	0.000	0.005	0.052	0.492	0.782	-	✓			
	C1	0.000	0.006	0.046							
	C2	0.000	0.005	0.042							
to conclude	B2	0.000	0.017	0.093	0.672	0.714	-	✓			
	C1	0.000	0.011	0.061							
	C2	0.000	0.006	0.041							
all in all	B2	0.000	0.007	0.060	6.342	0.042	B2 < C1		✓		
	C1	0.000	0.018	0.078			C1 = C2				
	C2	0.000	0.010	0.051			B2 = C2				
to sum up	B2	0.000	0.050	0.154	7.383	0.025	B2 > C1			✓	
	C1	0.000	0.015	0.073			C1 = C2				
	C2	0.000	0.021	0.073			B2 = C2				
	B2	0.000	0.052	0.156	9.051	0.011	B2 > C1				✓

In conclusion	C1	0.000	0.015	0.073			C1 < C2 B2 = C2				
	C2	0.000	0.028	0.083							

Logical connectives

Item	Level	Median	Mean	SD	X^2	p	Post-hoc summary	Pattern of change			
								no diff	up	down	down-up
also	B2	0.000	0.227	0.365	3.256	0.196	-	✓			
	C1	0.000	0.230	0.317							
	C2	0.273	0.232	0.297							
although	B2	0.000	0.071	0.202	0.017	0.991	-	✓			
	C1	0.000	0.051	0.138							
	C2	0.000	0.044	0.119							
besides	B2	0.000	0.014	0.094	2.202	0.333	-	✓			
	C1	0.000	0.014	0.081							
	C2	0.000	0.014	0.064							
in addition	B2	0.000	0.021	0.102	3.223	0.200	-	✓			
	C1	0.000	0.027	0.100							
	C2	0.000	0.020	0.071							
on the contrary	B2	0.000	0.003	0.042	5.168	0.075	-	✓			
	C1	0.000	0.005	0.041							
	C2	0.000	0.009	0.054							
on the other hand	B2	0.000	0.106	0.218	2.836	0.242	-	✓			
	C1	0.000	0.069	0.144							
	C2	0.000	0.053	0.110							
so	B2	0.052	0.079	0.105	2.988	0.224	-	✓			
	C1	0.075	0.081	0.086							
	C2	0.057	0.072	0.077							
as a result	B2	0.000	0.000	0.000	8.888	0.012	B2 = C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.001	0.021							
	C2	0.000	0.005	0.038							
consequently	B2	0.000	0.007	0.060	12.506	0.002	B2 = C1 C1 < C2 B2 < C2		✓		
	C1	0.000	0.010	0.065							
	C2	0.000	0.020	0.084							
even though	B2	0.000	0.022	0.106	9.605	0.008	B2 < C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.039	0.123							
	C2	0.000	0.035	0.099							
furthermore	B2	0.000	0.022	0.106	7.871	0.020	B2 = C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.029	0.105							
	C2	0.000	0.032	0.093							
however	B2	0.000	0.114	0.239	12.288	0.002	B2 = C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.137	0.236							
	C2	0.000	0.146	0.215							
moreover	B2	0.000	0.035	0.130	15.416	0.000	B2 = C1 C1 < C2 B2 < C2		✓		
	C1	0.000	0.029	0.101							
	C2	0.000	0.048	0.111							
nevertheless	B2	0.000	0.016	0.089	8.272	0.016	B2 = C1 C1 = C2 B2 < C2		✓		
	C1	0.000	0.024	0.094							
	C2	0.000	0.025	0.083							
since	B2	0.000	0.019	0.107	44.288	0.000	B2 < C1 C1 < C2		✓		
	C1	0.000	0.033	0.116							

	C2	0.000	0.066	0.138			B2 < C2				
therefore	B2	0.000	0.040	0.157	74.343	0.000	B2 = C1		✓		
	C1	0.000	0.026	0.102			C1 < C2				
	C2	0.000	0.105	0.186			B2 < C2				
though	B2	0.000	0.010	0.073	8.746	0.013	B2 ≤ C1		✓		
	C1	0.000	0.021	0.090			C1 = C2				
	C2	0.000	0.025	0.120			B2 < C2				
thus	B2	0.000	0.007	0.060	35.265	0.000	B2 = C1		✓		
	C1	0.000	0.014	0.076			C1 < C2				
	C2	0.000	0.037	0.104			B2 < C2				
whereas	B2	0.000	0.014	0.084	11.160	0.004	B2 = C1		✓		
	C1	0.000	0.010	0.058			C1 < C2				
	C2	0.000	0.022	0.077			B2 < C2				
while	B2	0.000	0.064	0.181	13.173	0.001	B2 = C1		✓		
	C1	0.000	0.044	0.125			C1 < C2				
	C2	0.000	0.076	0.155			B2 < C2				
yet	B2	0.000	0.003	0.042	25.387	0.000	B2 ≤ C1		✓		
	C1	0.000	0.013	0.073			C1 < C2				
	C2	0.000	0.028	0.094			B2 < C2				
because	B2	0.519	0.490	0.584	52.059	0.000	B2 = C1			✓	
	C1	0.357	0.512	0.583			C1 > C2				
	C2	0.000	0.232	0.353			B2 > C2				
but	B2	0.519	0.730	0.665	14.560	0.001	B2 = C1			✓	
	C1	0.714	0.645	0.480			C1 > C2				
	C2	0.546	0.496	0.387			B2 ≥ C2				
and	B2	2.595	2.687	1.352	53.399	0.000	B2 > C1				✓
	C1	2.143	2.308	1.147			C1 < C2				
	C2	3.006	3.012	1.208			B2 = C2				

Sequencing

Item	Level	Median	Mean	SD	X^2	p	Post-hoc summary	Pattern of change			
								no diff	up	down	-
finally	B2	0.000	0.034	0.109	5.219	0.074	-	✓			
	C1	0.000	0.018	0.078							
	C2	0.000	0.031	0.100							
firstly	B2	0.000	0.031	0.123	0.163	0.922	-	✓			
	C1	0.000	0.023	0.096							
	C2	0.000	0.015	0.063							
last	B2	0.000	0.011	0.025	1.474	0.479	-	✓			
	C1	0.000	0.019	0.056							
	C2	0.000	0.010	0.024							
lastly	B2	0.000	0.007	0.060	0.219	0.896	-	✓			
	C1	0.000	0.004	0.036							
	C2	0.000	0.003	0.027							
next	B2	0.000	0.011	0.044	3.465	0.177	-	✓			
	C1	0.000	0.027	0.095							
	C2	0.000	0.012	0.056							
secondly	B2	0.000	0.026	0.113	0.277	0.871	-	✓			
	C1	0.000	0.021	0.085							
	C2	0.000	0.015	0.061							
thirdly	B2	0.000	0.005	0.052	1.021	0.600	-	✓			

	C1	0.000	0.002	0.029							
	C2	0.000	0.001	0.016							
to start with	B2	0.000	0.000	0.000	2.805	0.246	-	✓			
	C1	0.000	0.002	0.029							
	C2	0.000	0.003	0.027							
first	B2	0.000	0.069	0.129	12.119	0.002	B2 = C1 C1 < C2 B2 = C2		✓		
	C1	0.000	0.027	0.052							
	C2	0.000	0.059	0.095							

Topic shifts

Item	Level	Median	Mean	SD	X^2	*p*	Post-hoc summary	Pattern of change			
								no diff	-	down	-
now	B2	0.000	0.020	0.064	3.067	0.216	-	✓			
	C1	0.000	0.011	0.028							
	C2	0.000	0.010	0.023							
well	B2	0.000	0.029	0.084	15.053	0.001	B2 > C1 C1 = C2 B2 > C2			✓	
	C1	0.000	0.026	0.053							
	C2	0.000	0.007	0.011							