Comparing automated analysis and human analysis of hedging in academic texts

Rachael Ruegg
Victoria University of Wellington, New Zealand

Pansa Prommas
Prince of Songkla University, Thailand

Hedging is a metadiscourse device employed by academic writers to manage knowledge claims and establish writer-reader interaction in written discourse. Research writing involves a balance of fact and a writer's personal evaluation and interpretation. This study compared automated analysis of hedging through Authorial Voice Analyzer (AVA) with a more traditional human analysis of hedging, to increase understanding of the relative strengths and weaknesses of the AVA versus human analysis of hedging in academic texts. An explanatory sequential mixed-methods design was used; quantitative analysis (Pearson correlation) was followed by qualitative analysis to understand the reasons for quantitative differences. AVA found a larger number of hedging items than the human analysis in the same academic writing corpus. However, qualitative analysis suggests that the AVA only considers frequency and does not take account of function. Since many hedging devices are multifunctional, AVA seems to overestimate the frequency of hedging by counting hedge markers as hedging even when they are used with propositional functions. Overall, automated analytic tools like AVA are useful for metadiscourse studies. However, unless used in combination with human analysis they are unlikely to effectively operate with multifunctional markers. The findings of this study offer validation of AVA and help raise awareness of how the tool can be used to evaluate hedging in academic texts.

Keywords: Metadiscourse, hedging, automated analysis, corpus analysis, academic writing

Email address for correspondence: rachaelruegg@gmail.com

Introduction

Metadiscourse is the components of a text that do not constitute propositional content. It consists of markers to guide the reader through the text as well as to create reader-writer interactions in the text. Metadiscourse has been divided into two categories based on its functions: textual and interpersonal metadiscourse (see Vande Kopple, 1985). Interpersonal metadiscourse is that which allows the writer of a text to interact with readers, to express their evaluations, thoughts, and feelings about the propositional content of the text (Hyland, 2019). In formal academic register, hedging is a very common type of interpersonal metadiscourse as use of hedging shows the writer's anticipation of alternative perspectives and constitutes an attempt to maintain politeness in interaction between writers and readers, especially when making arguments that do not align with previous work. Hedging markers include words such as 'may' and 'possibly', which show awareness that other interpretations are possible. Thus, hedging is one way of showing caution in academic writing. People who are learning to write for academic purposes are often taught to be cautious when making claims, and such caution is seen as an indicator of academic competence and of quality in academic writing. The importance of learning to hedge in order to meet the expectations of readers of academic writing results in a need to measure and evaluate the extent to which academic texts are adequately and appropriately hedged.

The need to measure and evaluate the extent to which texts meet certain expectations of readers, or the ways in which they fail to do so, entails the need to accurately analyse linguistic features of texts. Manual text analysis is incredibly time-consuming. Thus, many tools have been developed for automated analysis of texts. Preliminary examples include tools such as the grammar and spelling checker on word processing software. However, recently a large number of tools have been developed to analyse an ever-increasing range of linguistic features. For example, Coh-metrix (Graesser et al., 2011) automatically analyses texts for 108 different indices, and Tool for Automatic Analysis of Lexical Sophistication (TAALES) (Kyle & Crossley, 2015) automatically measures 484 different indices, all of which relate to vocabulary. Such tools have revolutionized linguistic analysis of texts. However, robust validation is necessary to ensure that such tools measure what they purport to measure and that the measurements they provide are accurate and meaningful in relation to each index measured.

Automated analysis was introduced in the assessment of TOEFL written texts (Attali, 2007) after it was found that the automated scores achieved through the use of e-rater were significantly predictive of human scores. However, as mentioned by Attali and Powers (2008), the features measured through e-rater are surface-level writing features, such as grammatical and lexical accuracy and complexity. While such surface-level features are an important aspect of the assessment of L2 writing and at lower levels of writing proficiency, measures related to register and tone are more likely to be important at higher writing proficiency levels and when assessing the quality of L1 writing. Quinlan et al. (2009) mentioned that future development of e-rater should focus on deeper rhetorical issues such as content and organization, as well as on increased accuracy in the surface-level language features already measured by e-rater. However, they did not mention another worthy direction of development: deepening the level of language measured by adding sociolinguistic features, such as register and tone. Banerjee et al. (2015) conducted a validation of the Examination for the Certificate of Proficiency in English (ECPE) using a mixed-methods design and found that authorial voice was an important aspect of academic writing that should be added to the test's rating scale. Moreover, Bax et al. (2019) found that the quantity of hedging in writing decreased significantly as L2 student proficiency levels increased. Thus, authorial voice would be an important addition to automated evaluation tools of both L1 and L2 academic writing to deepen the level at which language is measured to the sociolinguistic level.

The AVA (Yoon, n.d.) is an automatic analysis tool which is freely available online. It measures frequency of features that contribute to authorial voice in academic texts. AVA has been used for quantitative analysis in essays, but its strengths and weaknesses versus human analysis of hedging in advanced academic texts like thesis writing has not previously been explored. The purpose of this study is therefore to determine the strength of the relationship between automated analysis of hedging using AVA and human analysis of hedging in thesis writing. The results are expected to be useful in increasing the quality of academic writing, and in assessment of academic writing quality.

The research questions are as follows:

1. What are the frequencies of hedging markers using AVA and human analysis?

2. What is the relationship between AVA and human analysis?
3. What causes disparities between AVA and human analysis?

Literature review

Hedging

Within interpersonal metadiscourse, hedging is a rhetorical strategy that academic writers employ to manage their knowledge claims and establish writer-reader interactions in written discourse (Hyland, 2019). Research writing involves a balance of fact and a writer's personal evaluation and interpretation of research entities (e.g., findings). Explicit use of hedging therefore enables readers to distinguish between factive claims, based on factual certainty or accepted knowledge and non-factive claims, based on probability of propositional truth (Hyland, 1996).

When making non-factive or opinion-based claims in research writing, writers are at risk of attracting negative responses from skeptical scholars in their disciplinary communities. It is, therefore, essential for writers to ensure that their propositions are not overstated to gain acceptance from readers. In this sense, hedging is considered as caution used by writers to weaken the force of propositions which may be presented from a subjective and epistemic perspective (Hinkel, 2005). Hedged propositions can express anticipation of criticism and possible refutation that might happen due to insufficiency of evidence and literature to support the claims being made (Hyland, 1996).

According to Hyland (2019), not only non-factive but also factive claims presented from an objective or evidence-based perspective are often hedged, especially when such claims seem to challenge existing knowledge of researchers and scholars. Explicit use of hedges for factive and objective claims shows the writers' awareness of relationships between writers and readers in academic discourse. This also implies that the motivation for hedging in written discourse is not only based on the limitation of propositional accuracy and certainty but also pragmatic reasons.

Based on politeness strategies (see Brown & Levinson, 1978; Myers, 1989), writers are expected to show awareness of readers in terms of social power or status. Criticizing or arguing against existing theories and making assertive claims with no room for reader negotiation or choice may be considered impolite and face-threatening acts. To express arguments in an acceptable manner, hedging is a politeness strategy to mitigate imposition, avoid face-threatening acts, and thus build solidarity with readers (Hyland, 2019; Myers, 1989).

In addition, Hyland (2005) distinguishes three functions of hedges: precision, prudence, and politeness. First, hedges attest to the degree of precision or reliability of propositions (e.g., the writer's reasoning vs. certainty). Second, 'prudence' refers to writers' avoidance of personal responsibility for statements to protect their reputation and limit damage which may result from categorical commitments. Third, 'politeness' is to develop writer-reader relationships and solidarity by stating claims as alternative viewpoints and leaving them open to reader evaluation.

In sum, hedging markers are used to shield discourse actions from criticism and show readers the degree of uncertainty and approximation when there is not sufficient evidence or literature to support propositional content. Writers also hedge when they have doubts about an assertion to leave room for alternative viewpoints and thus encourage reader participation in the discourse. Common realizations of hedging are through modal verbs *(may, could)*, epistemic verbs *(appear, suggest)*, adverbs *(possibly, often)*, and adjectives *(likely, possible)* (Crismore et al., 1993; Hyland, 2019; Vande Kopple, 1985; Williams, 1981).

Automated analytic tools and corpus analysis

As mentioned above, automatic text analytic tools have revolutionized linguistic analysis of texts. Measures that would previously have taken months to analyse manually, can now be analysed in a matter of minutes. Although speed of analysis is an obvious advantage of automated analysis, there is a need for deeper understanding of the relationships between automated analysis and human analysis (Crossley & McNamara, 2012). This relationship needs to be examined independently for each linguistic feature. A study by Connor-Linton and Polio (2014) found disparities between automated analysis and human assessments of accuracy. The authors concluded that humans reached different evaluations of accuracy when the texts were in fact equal in terms of accuracy. However, as mentioned by Jarvis (2017), automated tools rely heavily on frequency, so that an automated analysis of accuracy would be based solely on the number of errors present. On the other hand, humans are able to consider not only the frequency of errors but also the relative gravity of different kinds of errors. This means that humans may rate a text that has more important errors as less accurate than one which has only minor errors, although the minor errors may be more frequent. It could be argued that the human assessments in Connor-Linton and Polio (2014) were not less accurate than the automated analysis but rather more nuanced, relying on subtle features of language that go beyond simple frequency.

For reasons such as this, the main conclusion of most researchers has been that automated analysis should complement human analysis rather than serving as an alternative to it (Bektik, 2019; Cotos, 2014; Cushing, 2017; Matthews & Wijeyewardene, 2018). In the development of automated analytic tools, human judgements could be used as the foundation upon which to build an automated analytic tool (Jarvis, 2017). On the other hand, for researchers the most effective method of combining etic and emic perspectives depends on the kind of automated tool utilized and the purpose of the analysis. When using a tool which not only measures but also tags instances of the items under analysis, human analysis of the output can follow automated analysis. Tools which are relatively less automated allow for greater human judgement in the sense that the user can identify the actual features whose frequency should be measured by the tool in addition to subsequently evaluating the individual instances of each feature in the output. Overall, the most automated analyses offer the most etic perspective, while more simple tools that allow human judgement in what to count and how to count it offer a perspective which balances the advantages of automated analysis with nuanced human understanding of language and meaning.

Automated analysis of hedging

We identified three tools for the automated analysis of hedging in written texts, which will each be discussed here, as well as providing an overview of what has been found in relation to the frequency of hedging in academic texts through the use of such tools.

Multidimensional analysis tagger (MAT) (Nini, 2015) measures hedging, but only includes informal hedge markers. The features tagged and analysed by the programme are: "maybe, at about, something like, more or less, sort of, kind of" (p. 26). These features would be useful for measuring hedging in informal texts, but this tool is unable to measure hedging in formal academic prose. MAT also measures downtoners which can be used for hedging. The frequency of the following words is analysed by the programme: “almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat” (p. 26). Although many of these words are likely to be infrequent in academic prose, two in particular are fairly commonly used for hedging in this context: almost and somewhat. Overall, while some hedging devices that are relevant to academic register are measured by this tool, many are omitted from analysis, making the tool ineffective for measuring hedging in formal academic writing.

Hedge-O-Matic (Omizo & Hart-Davidson, 2016) is another tool for automated analysis of hedging. It analyses texts based on a wide range of hedging features, therefore being appropriate for use with either informal or formal written texts. However, in contrast to most automatic text analytic tools, Hedge-O-Matic measures the proportion of sentences that contain hedging rather than the overall frequency of hedging in a text. This would seem to be a less precise way of measuring hedging as some sentences may contain multiple instances of hedging, while others may not contain any. Compounding this is the effect of different sentence lengths. Analysis at the sentence level is difficult to reproduce as most tools do not analyse texts at the sentence level. Although this very different method of analysis merits further consideration, it would appear to lead to less precise measurement of hedging as measuring the percentage of sentences including hedging is less precise than knowing the exact number of times hedging is used in a text.

AVA (Yoon, n.d.) was designed specifically for the analysis of formal academic written texts. Yoon (2017) validated the AVA as a whole, by comparing overall AVA scores with holistic voice strength scores assigned by expert human raters. In addition, the AVA scores were also compared to overall essay quality scores assigned by expert human raters. While the results of Yoon (2017) contribute to our understanding of the impact of frequency of hedge markers on human assessments of authorial voice and of essay quality, the study did not consider the function of each marker in context. The hedge markers in AVA were identified based on previous literature, but every occurrence of each marker is counted by AVA and included in the output related to the number of hedge tokens and hedge types. No further analysis is conducted by the programme to determine the likelihood of each occurrence serving a hedging function. Thus, the output related to the number of hedge types and the number of hedge tokens likely overestimates the actual frequency of hedging in texts.

Interestingly, Yoon (2017) found that higher frequency of boosters, attitude markers and self-mentions were significantly predictive of holistic voice scores assigned by human raters, whereas frequency of hedging was not. On the other hand, frequency of hedging was the only index measured by AVA that predicted higher overall essay quality scores. Another similar study (Prommas, 2020) looked at the ability of different categories of metadiscourse to predict writing quality scores in postgraduate writing and found hedging to be significantly predictive of writing quality scores across both disciplines analysed. The results of both of these studies show the unique importance of hedging compared to other kinds of metadiscourse for quality formal academic writing.

Another study by Yoon and Römer (2020) used data from AVA to analyse stance and engagement markers in academic writing of final year undergraduate students and postgraduate students. The study analysed hedges, boosters, and attitude markers. In addition to automated analysis, the researchers qualitatively hand coded the hedging in 20 texts to validate the automated analysis of hedging conducted by AVA. They found a reliability of 0.921 between the qualitative hand coding and the automated analysis. However, only 20 texts were qualitatively analysed, and it is not clear whether this result would be maintained with a larger sample.

Several previous studies have looked at the frequency of hedging in academic writing in different contexts. Of relevance here are studies that have focussed on hedging in social science disciplines. Hu and Cao (2015) found between 57 and 121 hedges per 10,000 words across applied linguistics, education and psychology research papers, while Chen and Zhang (2017) found 184 hedges per 10,000 words in applied linguistics research articles. In a diachronic study of hedging from 1965 to 2015, Hyland and Jiang (2018) found the number of hedging markers per 10,000 words to be decreasing through time in both applied linguistics (from 201 to 129) and sociology (from 188 to 149). Thus, an appropriate level of hedging for student writers could range anywhere between around 60 and around 200 hedges per 10,000 words.

Overall, despite a number of studies on hedging in academic writing, there is still little clarity about how much hedging could be considered appropriate in different academic writing contexts, such as thesis writing. However, underuse or overuse of hedging in academic texts is likely to reduce the effectiveness of such texts, thus further research is needed to establish reliable parameters for effective hedging use. Such parameters would be an important aspect of effective assessment of academic text quality.

Methods

The current study compares automated analysis of hedging, as measured by AVA, with a more traditional human analysis of hedging using Wordsmith (Scott, 2016). Once a list of items is identified and entered into Wordsmith, it can identify and tag instances of those items, which can then be considered individually to determine their suitability for inclusion. One hundred and sixteen masters' theses in two disciplines will be analysed through both methods. While numerous studies have focused on analysis of research articles written by professional writers, this study has extended the focus to analysis of master's theses written by novice research writers. Although both tasks are research-oriented writing genres, the latter is produced for the purpose of disciplinary knowledge assessment by specific target readers (i.e., supervisors and examiners). Hedging markers play a very important role in discussion and conclusion chapters to make cautious claims and establish prudent authorial voice when interpreting research results and providing justifications for findings. Therefore, discussion and conclusion chapters were the focus of this study. The aim of the study is to increase understanding of the relative strengths and weaknesses of the AVA versus human analysis of hedging in thesis writing.

Research orientation

This research follows an explanatory sequential mixed-methods design (Creswell, 2017). Quantitative analysis of all the texts was conducted first, through the two analytical methods in focus. Following this, qualitative analysis of a smaller number of texts was carried out to help explain the quantitative results. Since the qualitative analysis was intended to focus on differences between the two analytical methods, purposive extreme case sampling was employed to select cases for the qualitative phase of the research (Dörnyei, 2007).

Corpus compilation

This study is based on a corpus of 116 discussion and conclusion chapters of master's theses written in English by New Zealand and Thai postgraduates in the disciplines of English language teaching and business administration. This is a convenience sample, as the corpus was collected for a previous study and reused here. New Zealand theses were collected from six universities in New Zealand, and Thai theses were collected from two universities in Thailand. Only theses for which consent had been given by the New Zealand authors and Thai universities were included in this study. The whole corpus contained 599,599 words, with an average length of 5,169 words per thesis.

Analysis of hedging markers

The human analysis was conducted first. Taking a top-down approach, a hedging marker list used in the human analysis in this study was developed based on three sources. Hyland's (2005) list of metadiscourse markers was used as a starting point for creating a pre-identified marker list. Using the pre-identified marker list makes recognizing potential items easier for coders dealing with a large number of items in lengthy texts. To expand Hyland's list, we added the potential hedging markers found in the thesis corpus and the Academic Word List (AWL) (Coxhead, 2000). To do so, the tool 'WordList' in Wordsmith (Scott, 2016) was used to generate a list of common words extracted from the thesis corpus. The program generated a list of more than 10,000 common words used in the corpus and from this list some potential hedging markers which are not present on Hyland's list were identified (e.g., *incline, imply, basically*). The AWL was also used in this study because words in the AWL are from academic texts at tertiary level (e.g., academic book chapters, journal articles) across a wide range of disciplines (e.g., arts, commerce). The level of texts and disciplines are therefore suitable for this study. In the human analysis, a separate automatic search of each hedging marker was done using Wordsmith (Scott, 2016). Each concordance line was examined manually to eliminate any items that did not express the hedging function. To ensure the coding reliability of this study, the researchers set a coding scheme and noted reasons for the inclusion or exclusion of ambiguous hedging markers. Intra-coding reliability was obtained by recoding the data three months after the first round. The agreement rate between the two sets of results was 95%. Disagreed cases were resolved through discussion and the guidance of an expert on metadiscourse. Table 1 shows the coding scheme samples of *argue* and *frequently*.

Table 1. Examples of human coding of metadiscourse

Argue	**X (external source) argue ...... = not hedging**
Included	*While minor discursive tussles provided an occasional stimulus for renegotiation (cf Miller, 2004; Park, 2007), I would argue that this particular context appears to have been an extremely valuable rehearsal space.* [NZ-ELT-05]
Included	*It could be argued that increased centralisation would be detrimental to the federation as affiliates and non-affiliates may baulk at the loss of operational autonomy and increased REPs NZ dependency.* [NZ-BA-13]
Excluded	*Haston and Russell (2012) argued that exposure to real life situations might prepare and strengthen teachers pledge to teaching if they feel successful during genuine learning situations.* [NZ-ELT-11]
Frequently	***-The item shows uncertainty (not 100% commitment) = hedging*** ***-The item refers to actual facts/ frequency in the studies = not hedging***
Included	*Their self-assessments frequently failed to distinguish between evaluations of the learning methods, learning outcomes, or their match to their language needs.* [NZ-ELT-10]
Excluded	*Group 1 (the highest proficiency group) used all six strategy categories more frequently than the other 3 groups, while Group 2 used them more frequently than Group 3 and so on.* [TH-ELT-12]

As the aim of this study was to compare the results of AVA (Yoon, n.d.) and human analysis of hedging, facilitated by Wordsmith (the traditional program widely used for text analysis). Following the human analysis, the same corpus was analysed by AVA (although for this study only the results related to hedging were used). The AVA results in this study were the number of hedge tokens found by AVA in each text. The results of the two analyses were then explored graphically using a scatterplot and numerically using Pearson correlation.

Qualitative analysis of discrepancies

Following the quantitative analysis, qualitative analysis was employed to understand factors that led to different results between the number of hedges found by AVA and human evaluation. The first stage of the qualitative analysis involved comparing the list of hedge items included in AVA with the final list of hedges developed for the human analysis from both the literature and the corpus. These two lists were examined to evaluate the effect of differences between the two lists on the quantitative results.

Following this, extreme case sampling was used to identify four texts which exhibited very different results between the two analyses: Two which were found to have more hedges in the human analysis and two which were found to have more hedges in the AVA analysis. All items on the two lists of hedges used in the human analysis and AVA were examined in context in these four texts to determine the impact of the two different methods of analysis. Although AVA provides raw frequency of hedges in each text, it does not provide frequency data for each hedge item. We were therefore unable to compare frequencies between AVA and Wordsmith. Thus, we searched for and evaluated items in context that may cause disparities between automated analysis and human analysis methods based on Wordsmith frequency data of each item before and after human analysis.

Results

Quantitative results

Descriptive statistics for both analyses of hedging can be seen in Table 2. Following the computation of descriptive statistics, Pearson correlation was run and the correlation between the two methods of measurement was found to be: r (115) = 0.855, p = <.001, indicating a significant correlation between the two analytical methods and a large effect size.

Table 2. Descriptive statistics for frequency analysis

Method	Mean	SD	Minimum	Maximum
AVA	82.3276	71.0742	7	406
Human analysis	66	53.0035	10	339

Similarly, the scatterplot, which can be seen in Figure 1, shows a linear relationship between the two analytical methods. However, a paired samples t-test between the two methods found a large and statistically significant difference between the results (t (115) = 4.672, p <.001). This indicates that, while there was a significant correlation between the two measures and a large effect size, there were nevertheless also significant differences between the analytical methods. This significant difference suggested that it would be worthwhile to conduct a qualitative analysis to gain a deeper understanding of the differences.

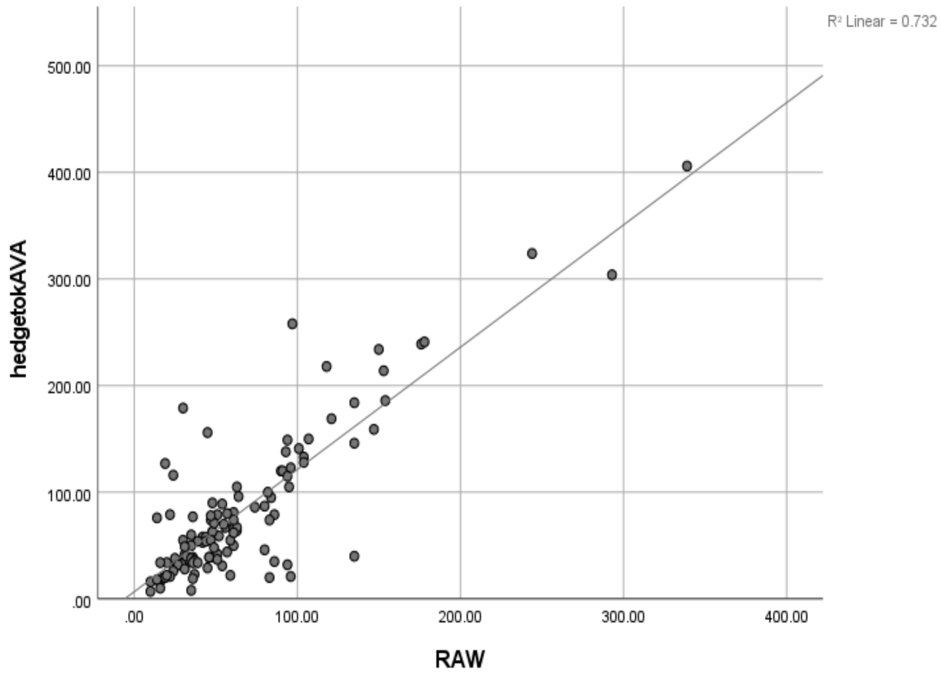

Figure 1. Scatterplot of relationship between analytical methods

A more detailed impression can be achieved by comparing the raw frequency of hedge items initially found by Wordsmith with the final number of hedge items determined after human analysis and comparing those two figures to the frequency found in AVA. These three frequency figures can be seen in Table 3.

Table 3. Frequency results of Wordsmith with and without human analysis and AVA

	Wordsmith raw	Wordsmith with human analysis	AVA
NZ-ELT	4370	2764	3380
NZ-BA	2560	1842	2295
TH-ELT	2203	1560	1828
TH-BA	2101	1525	1547
Total	11234	7691	9050

This comparison shows that the raw frequency of hedge tokens as measured by AVA differs quite considerably from both the raw frequency of hedge items found through Wordsmith and the final number of hedge tokens determined through subsequent human analysis. While AVA found substantially fewer tokens than raw Wordsmith output, it found substantially more than subsequent human analysis of Wordsmith output.

Qualitative results

Identification of hedges by automated analytic tools like AVA rely more heavily on frequency regardless of meaning. On the other hand, human analysis usually involves an interpretation of both form and meaning. Hedges are categorized as interpersonal metadiscourse markers that may perform more than one function and thus could not be identified by form only (Hyland, 2005). Careful interpretation of their functional meanings to identify whether potential items are hedges in the category of interpersonal metadiscourse is a crucial part of metadiscourse analysis which automated tools may fail to achieve. This section identifies discrepancies between the items identified during the human analysis and automated analysis using AVA in order to determine the validity of automated analysis of hedging using this tool.

Out of 108 pre-identified markers used in the human analysis, 77 items (about 70%) overlapped with those on the AVA list. Out of 172 pre-identified markers on the AVA list, only 59 items (about 34%) were ones which had been included in the human analysis. About 113 items (66%) were omitted from the human analysis. Despite more types of hedges in the AVA analysis and somewhat different items between the two methods, the quantitative results in the previous section showed similarity in overall frequency and a significant positive correlation between the automated analysis and human analysis methods.

The first step in the qualitative analysis involved adding the different items from the AVA list (113 items) to the human analysis of 116 theses to investigate the impact of those items on overall hedging frequency. The results suggest that those additional items were not popular as they were not frequently found in the thesis corpus (73 tokens or 0.8% of all occurrences in AVA). This means that even though the two methods include some different items, the additional items included on the AVA list are unlikely to have had a significant effect on the frequency results or to have caused disparities between the two methods.

To take a closer look at the items that overlap between AVA and human analysis (77 items), four extreme cases with the greatest differences in frequency of hedges between AVA and human analysis were selected and hand coded to further investigate the disparities found between the two analytical methods. To recall the notion of hedges, high frequency hedges can be grouped into four grammatical categories expressing the writer's evaluation of possibility, uncertainty, and necessity of propositions: modal verbs (*may, would, could*), verbs (*suggest, indicate, tend to*), adjectives (*likely, possible),* and adverbs (*mostly, frequently, often*). In the analysis of the four theses, the most frequent method of hedging was through use of modal verbs (*may, should, could, might*). The result confirms the initial analysis of the corpus of 116 theses, indicating modal verbs as the most salient strategy to soften arguments in thesis writing. As shown in Table 4, the top five markers (*may, should, could, frequently,* and *might*) account for more than half of all occurrences of hedges (54.8%) in the four theses. These highly frequent hedging markers were therefore the first set of target words analysed, followed by other suspicious markers.

Table 4. Frequent hedging types in the four theses

Frequent markers used in the 4 theses	Raw frequency before human analysis (N = 347)	%	Frequency after human analysis (N = 272)	%
may	52	15.0	52	19.1
should	43	12.4	39	14.3
<u>could</u>	44	12.7	21	7.7
<u>frequently</u>	35	10.1	21	7.7
might	16	4.6	16	5.9
Total	190	54.8	149	54.8

Notes. N = Total occurrences of hedges in the four theses.
Underlined markers show considerable frequency differences between the two methods.

May/ Should/ Could/ Might

As shown in Table 4, a comparison of frequencies of *may, should, could,* and *might* before and after human analysis indicates similarity between automated analysis and human analysis methods in quantifying the frequencies of the most frequent modal verbs. Particularly, the results of *may, should,* and *might* between the two methods are strikingly similar. A substantial difference was found only in the result of *could* (44 tokens found by Wordsmith without human analysis, 21 tokens remaining after human analysis). One possible reason for the difference is that, without human interpretation, it might not be feasible for automated tools to distinguish functions of modal verbs, especially *could*, as they sometimes appear in discourse as part of ideational content (i.e., expressing ability in the real-world context) rather than being hedging devices. Meanwhile, *may* and *might* literally and explicitly express tentativeness and epistemic possibility of propositions.

Excerpts 1 and 2 illustrate the role of *could* as an ideational marker expressing the (in)ability of participants to answer questions and use many words in writing, respectively. By contrast, Excerpts 3-5 show *could* as an interactional metadiscourse marker expressing the plausible justification of the research findings and tentative recommendations for future studies.

> (1) They thought that if a teacher ask them back about word meanings, they would be embarrassed because they could not answer the teacher's questions. [TH-ELT-07]

> (2) Moreover, the table also indicates that the subject's fluency greatly improved as she could produce 80 more words in her final journal entry compared to her first. [TH-ELT-08]

> (3) This could be attributed to their positive attitudes toward the use of journal writing with peer feedback, reinforcing a positive attitude toward writing in English. [TH-ELT-08]

> (4) One reasonable explanation could be that the students may not be aware of the importance of high-frequency vocabulary. [TH-ELT-07]

> (5) ...further study could be conducted by combining qualitative approaches, such as in-depth interviews or case studies, to complement the quantitative research methodology and obtain more comprehensive results... [TH-ELT-30]

Frequently

Evaluation of *frequently* as an ideational or interactional device is also a case of context dependence. Identifying *frequently*, as well as other adverbs of frequency, in isolation from its function in context might not lead to valid results. It is likely that automated analytic tools may quantify tokens of *frequently* based on its form only. As in Excerpts 6 and 7, use of *frequently* is not related to the interactional metadiscourse function of hedging; expressing epistemic stance of the writer (e.g., possibility, approximation, prediction) or reflecting the writer's commitment to the validity of the arguments. Rather, the following excerpts give factual information of frequency rather than denoting the writer's evaluation of the message.

> (6) *Hulstijn et al. (1996) asserts that EFL learners who frequently look up word meanings in a dictionary had better retention and retrieval of word meanings.* [TH-ELT-07]

> (7) *To discover and consolidate word meanings, Social Strategies were the least frequently use (57%) by the research subjects.* [TH-ELT-07]

Despite some occurrences of *often* in the corpus, the item was not identified as a hedging item when analysed in context. Similar to the case of *frequently,* it is possible that the item might be counted as a hedging marker by the AVA based on its form despite not serving a hedging function in the corpus texts.

> (8) *It is possible that they encountered the high-frequency vocabulary very often in a wide range of academic texts so they may assume that word frequency will provide them a basic vocabulary knowledge to understand English texts.* [TH-ELT-07]

> (9) *These words are useful to learn because they appear very often in academic texts and they equip learners with an adequate reading comprehension ability of a wide range of academic texts (Coxhead, 2000).* [TH-ELT-07]

Based on the notion that metadiscourse does not include ideational propositions outside the current text being discussed, the word *often* in Excerpts 8 and 9 simply functions as an adverb of frequency in the real-world proposition that the frequent *(often)* exposure to high-frequency vocabulary in a wide range of academic texts may affect the likelihood of comprehension.

Other hedging items that appear on the AVA list only

I think/I believe. In fact, *I think* and *I believe* can be identified as two different kinds of metadiscourse; either as hedging markers or attitude markers, depending on what kind of proposition follows the item. It can be inferred to be an expression of hedging when the proposition after *I think* or *I believe* is tentative and co-occurs with other hedging markers indicating the degree of uncertainty, such as epistemic modal verbs. The purpose of using *I think* or *I believe* is to indicate to readers that the proposition is the writer's own opinion or belief without their being concrete evidence to support the proposition. However, these items are unlikely to be found in research writing, given the fact that the research writing genre requires writers to rely more on evidence (collected data and statistical results) and literature rather than writers' assumptions. However, research writing also leaves room for writers to justify what they discover, so it is also possible that the markers might be used to express possible justifications of findings. From this point, it might be difficult for automated analytic tools to distinguish whether the items denote the writers' voice construction (possible justification) in the text or ideational content stated by someone else.

In the corpus analysed in this study, there are occurrences of *I think* and *I believe*. However, human analysis indicates that none of the occurrences of *I think* or *I believe* found in the thesis corpus serve hedging purposes. They are just propositions that were mentioned by research participants. Indeed, *I think* and *I believe* are not likely to serve hedging functions often in formal academic writing as they are not often considered appropriate to that register. Therefore, *I think* and *I believe* here do not refer to thoughts of the writers but quotes from their research participants, as in the excerpts below.

> (10) *So, my mother is the best woman who we should admire. All of her life she gives it to me. Although sometime we have a problem, she will solve it for us. I think she is the best person who we much take care nicely.* [TH-ELT-08]

> (11) *... so I realise someone I admire should be "my mother" I think everyone maybe think same with me. so we should spend the time to take care her, help her to do something and don't forget to say "Love" to her everyday. I believe everyone will be happy especially my mother who is the person which I admire.* [TH-ELT-08]

Hedging items that do not appear on the AVA list

Indicate/Claim. In human analysis, the reporting verbs *indicate (323 tokens)* and *claim (21 tokens)* were found to be high frequency items in the thesis corpus. However, it is interesting that *indicate* and *claim* which appear on Hyland's pre-identified marker list are not included on the AVA list of hedges. Another point is that those items can perform as hedges (Excerpts 12-13), evidentials (Excerpt 14), or even frame markers (Excerpt 18), so it is doubtful whether the AVA could accurately distinguish the functions of these multifunctional markers if they were included on the list.

> (12) *Findings claim that Indians complain that they have to wait for a long time while they are tired and hungry.* [THA-BA-17]

> (13) *The data indicated that factors that influence learners' idiosyncratic choices and decisions in turn affected the way the learners used their out-of-class language learning opportunities.* [NZ-ELT-10]

> (14) *Timperley et al. (2007) claim that school management teams are responsible for the professional development of staff; a more consistent approach to the induction of overseas teachers might help to decrease teacher turnover and its associated problems.* [NZ-ELT-14]

To further investigate multifunctional reporting verbs from the case of *suggest,* which is included on the AVA list and also in human analysis, this item showed a large number of occurrences (more than 800 tokens in the whole thesis corpus). However, based on human analysis, not all the tokens of *suggest* express the hedging function (about 500 tokens). In Excerpts 15-17, the role of *suggest* is to make a reference to external sources such as previous studies. Therefore, *suggest* is more likely to be an example of an evidential rather than hedging the writers' own discourse. In other words, the original suggestions are given by other researchers (e.g., Nation, 2009 and Holt, 2001), not the writers of the texts being analyzed.

> (15) *<u>Nation (1990), however, suggests that</u> it is worth spending time on in class to acquire these words for a larger vocabulary size.* [TH-ELT-07]
>
> (16) *<u>As suggested by Norris Holt (2001),</u> this type of motivation seems to be a powerful force influencing success in learning the language in the particular context in which language learners have very little substantial contact with the target language members outside of the classroom.* [TH-ELT-30]
>
> (17) *The result is in line with the study of Zhang (2009), Mochizuki (1999), and Nacera (2010) who found that EFL learners studying in a tertiary level used vocabulary learning strategies at a moderate level... Moreover, <u>these studies also suggest that</u> the EFL learners' frequency of use of vocabulary learning strategies seemed not enough for them to successfully learn vocabulary.* [TH-ELT-07]

Unlike the above cases, the word *suggest* in Excerpt 18 is used as a frame marker to provide readers a preview of what will happen in the following section. This is a metadiscursive function, but not an example of hedging.

> (18) *This chapter presents the summary of the study, denotes pedagogical implications, and <u>suggests</u> some recommendations for further studies.* [TH-ELT-08]

To identify this multifunctional reporting verb as a hedging item, the item should express, for example, tentativeness of claims being made. Based on the human analysis of the whole thesis corpus, a salient hedging pattern is collocating epistemic verbs *(suggest, indicate, claim)* with an inanimate subject (e.g., research itself, results, analyses), as exemplified in Excerpts 19-20.

> (19) *The findings also <u>suggested</u> that a potential source of the English teachers' lack great success in pragmatic competence was their linguistic deficiency.* [TH-ELT-03]
>
> (20) *The analysis <u>suggests</u> that attitudes towards English language played an important role in the subjects' success in learning vocabulary.* [TH-ELT-13]

From the analysis of *suggest*, it can be assumed that other reporting verbs with more than one function like *indicate* and *claim* might be items whose functions cannot be differentiated by automated analytic tools and which therefore still require human analysis.

Overall, despite the analysis of hedging through AVA relying on a different selection of hedging markers from the human analysis conducted for this research, the quantitative results show a great deal of similarity between the frequency of hedging found through the two methods. However, in-depth qualitative analysis indicates some notable differences between the frequency results found through the two methods for some specific items. These differences will be discussed in the next section.

Discussion

Results of the quantitative analysis indicate that AVA found approximately 25% more hedging devices in the corpus than the human analysis. This might lead one to believe that the AVA is more effective at finding hedging devices than a human. However, the subsequent qualitative analysis of four theses provided several alternative explanations for the different results, which will be discussed here.

Firstly, it is apparent that automated analysis, as conducted by AVA, considers occurrences of specific hedging items, but is not effective in taking account of the function of each item in its specific context. Analysing items for function in addition to form is important in the analysis of hedging because many hedging devices can be used to fulfil both hedging functions and real-world propositional functions, for example, *could* and *frequently* in this research. This suggests that for the automated analysis of any items which can be used for both hedging and non-hedging functions, it is essential to equip automated tools with more precise coding features to accurately identify the actual function at play each time such a marker is used.

Moreover, since many hedging devices are multifunctional in nature, this may be difficult to achieve in practice. The absence of a feasible way to programme automated tools to code items more precisely does not mean that automated analysis is not viable, but rather that it should be used in combination with human analysis, as suggested by

Bektik (2019), Cotos (2014), Cushing (2017), and Matthews and Wijeyewardene (2018). In this respect, the findings in this study may be of help to analysts in narrowing the scope of human analysis. In particular, some items (such as *may* and *might*) may be relatively unifunctional and thus less problematic to analyse with an automated tool such as AVA. On the other hand, particular items (i.e., *could, frequently, often, I think, I believe, suggest, indicate, claim*) would seem to benefit from the addition of human analysis of function after using the automated tool.

In addition to differences in the function of items analysed through the two different methods, the list of hedging items also differed between the human analysis and AVA. The list of items used in the qualitative analysis was created based on the list provided by Hyland (2005), in addition to the identification of additional items found to be frequent in the corpus, present in the AWL (Coxhead, 2000), and serving a hedging function. Once again, the list of items in AVA was larger than that used in the human analysis (approximately 34% more items), which again could be considered to be a strength of AVA in the analysis of hedging. Nevertheless, there were hedging markers commonly used in formal academic texts like theses absent from the AVA list (such as *indicate*, *claim* and *largely*), whereas hedging markers included in the AVA list which were not present in the list used for the human analysis were largely items that were not commonly used with hedging functions in the thesis writing corpus analysed here. This suggests that AVA is likely to overestimate the frequency of hedging in theses. Further research is suggested to shed more light on the frequency of those items in other corpora, in order to understand their relative usefulness for inclusion in lists of hedging items.

Conclusion

There are some limitations, which need to be taken into account when considering the results of this study. Firstly, the analysis conducted here only focused on theses, and only included theses from two disciplines and two countries. Analysis of other academic text types, or theses from different disciplinary or geographical contexts may lead to different results. Moreover, the lack of frequency data for each hedge type in AVA limited the extent of both quantitative and qualitative analysis that could be conducted. Despite these limitations, this study identified hedge types that could be effectively measured by analytic tools such as AVA, as well as some that require the addition of human analysis. These results will be helpful for text analytic tool developers, in developing more precise tools, as well as for those who use such tools for assessment of text quality or for research purposes, in determining where to focus subsequent human analysis.

Many automated analytical tools used in standardized tests of L2 writing, such as e-rater, only measure surface-level language features such as grammatical and lexical accuracy and complexity. There is a need for further research to determine the relative predictive ability of sociolinguistic measures such as measures of hedging in standardized language proficiency tests in addition to efforts to measure such features more accurately in automated analytic evaluation of writing quality.

Overall, automated analytic tools like AVA are useful for metadiscourse studies, offering a rough estimation of the frequency of different kinds of linguistic features, such as hedging. In addition, if a more accurate measurement is required, such tools can be used in eliciting potential target markers and estimating marker frequency. However, the results of analysis through automated tools still needs to be processed through human analysis given that the tools are not effective in identifying the function of each item in a given context and are therefore unlikely to effectively operate with multifunctional markers.

Author disclosures

The authors do not have any conflicts of interest that they are aware of.

Rachael Ruegg conceptualized the research project, designed the research methodology and contributed to writing, reviewing and editing of the article. Pansa Prommas collected the data for this project, carried out the qualitative data analysis, and contributed to writing, reviewing and editing of the article

ORCID iDs

Rachael Ruegg https://orcid.org/0000-0003-2437-3007

Pansa Prommas https://orcid.org/0000-0001-6450-5970

References

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series, 2007*(1), i-22. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y., & Powers, D. (2008). A developmental writing scale. *ETS Research Report Series, 2008*(1), i-59. https://doi.org/10.1002/j.2333-8504.2008.tb02105.x

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5-19. https://doi.org/10.1016/j.asw.2015.07.001

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

Bektik, D. (2019). Issues and challenges for implementing writing analytics at higher education. In *Utilizing Learning Analytics to Support Study Success (pp. 143-155)*. Springer.

Brown, P., & Levinson, B. (1978). Universals of Language: Politeness Phenomena, in E. Goody (Ed.), *Questions and Politeness.* Cambridge University Press.

Chen, C. & Zhang, L. (2017). An intercultural analysis of the use of hedging by Chinese and Anglophone academic English writers. *Applied Linguistics Review, 8*(1), 1-34. https://doi.org/10.1515/applirev-2016-2009

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, (26), 1-9. https://doi.org/10.1016/j.jslw.2014.09.002

Cotos, E. (2014). Automated writing evaluation. In *Genre-Based Automated Writing Evaluation for L2 Research Writing (pp. 40-64)*. Palgrave Macmillan.

Coxhead, A. (2000). A new academic word list. TESOL quarterly, 34(2), 213-238.

Creswell, J. W. (2017). *Designing and conducting mixed methods research.* SAGE Publications.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in Persuasive Writing. *Written Communication*, *10*(1), 39–71. https://doi.org/10.1177/0741088393010001002

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Cushing, S. (2017). Corpus linguistics in language testing research. *Language Testing, 24(4)*, 441-449. https://doi.org/10.1177/0265532217713044

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford University Press.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223-234.

Hinkel, E. (2005). Hedging, inflating, and persuading in L2 academic writing. *Applied Language Learning, 15*(1/2), 29-53.

Hu, G. & Cao, F. (2015). Disciplinary and paradigmatic influences on interactional metadiscourse in research articles. *Journal of English for Specific Purposes, 39, 12-25.* https://doi.org/10.1016/j.esp.2015.03.002

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, *17*(4), 433–454. https://doi.org/10.1093/applin/17.4.433

Hyland, K. (2005). Prudence, precision, and politeness: Hedges in academic writing. *Quaderns de Filologia-Estudis Lingüístics*, *10*, 99–112.

Hyland, K. (2019). *Metadiscourse: Exploring interaction in writing*. Continuum.

Hyland, K. & Jiang, F. (2018). "In this paper we suggest": Changing patterns of disciplinary metadiscourse. *Journal of English for Specific Purposes, 51, 18-30.* https://doi.org/10.1016/j.esp.2018.02.001

Jarvis, S. (2017). Grounding lexical diversity in human judgements. *Language Testing, 34(4),* 537-553. https://doi.org/10.1177/0265532217710632

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Matthews, J. & Wijeyewardene, I. (2018). Exploring relationships between automated and human evaluations of texts. *Language Learning and Technology, 22*(3), 143-158.

Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics, 10*(1), 1-35. http://dx.doi.org/10.1093/applin/10.1.1

Nini, A. (2015). *Multidimensional analysis tagger (version 1.3.1).* Retrieved from https://sites.google.com/site/multidimensionaltagger/

Omizo, R. & Hart-Davidson, B. (2016). *Hedge-o-matic.* Retrieved from http://hedgeomatic.cal.msu.edu/hedgeomatic/

Prommas, P. (2020). *Metadiscourse in postgraduate writing.* Te Herenga Waka—Victoria University of Wellington.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series, 2009*(1), i-35. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

Scott, M. (2016). *WordSmith Tools version 7.* Lexical Analysis Software.

Vande Kopple, W. J. (1985). Some Exploratory Discourse on Metadiscourse. *College Composition and Communication*, *36*(1), 82–93. https://doi.org/10.2307/357609

Williams, J. M. (1981). *Style: Ten lessons in clarity and grace.* Scott Foresman and Company.

Yoon, H. (n.d.) *Authorial Voice Analyzer.* Retrieved 3rd August 2021 from: https://hyungjo.weebly.com/ava.html

Yoon, H. J. (2017). Textual voice elements and voice strength in EFL argumentative writing. *Assessing Writing*, *32*, 72-84. https://doi.org/10.1016/j.asw.2017.02.002

Yoon, H. J., & Römer, U. (2020). Quantifying disciplinary voices: An automated approach to interactional metadiscourse in successful student writing. *Written Communication*, *37*(2), 208-244. https://doi.org/10.1177/0741088319898672