A Cross-Cultural Corpus Study of the Use of Hedging Markers and Dogmatism in Postgraduate Writing of Native and Non-native Speakers of English

Rawy A. Thabet

Abstract This study investigates the frequency of hedged propositions in academic writing, which are produced by both native (NSs) and non-native speakers (NNSs). To this end, two corpora, which represent native and non-native writings respectively, are compiled and investigated using contrastive interlanguage analysis (CIA). This computer-aided investigation, which involves comparing quantitative and qualitative data, is adopted to identify what the most frequent hedging markers, used by native and non-native writers, are, and whether there is any significant difference between the frequencies of these markers in both writings. This research is an attempt to fill a gap in literature, as there is a paucity of studies written on corpus analysis in the Middle East. The findings suggest that non-native speakers underuse hedges and the quality of these hedges is usually not so high as those of the native speakers.

Keywords Corpus analysis · Native speakers · Non-native speakers
Hedges · Modality · Overuse · Underuse

1 Introduction

One of the main problems that is faced by non-native speakers is the inability to express their stance or point of view without being dogmatic/hyperbolic. Scarcella and Brunak [1] admitted that the Arab (as an example of non-native speakers) learners lack the competence of using hedges. However, when the literature of modality was reviewed, it was found that this incompetence is not confined to Arabs but rather it is a common feature among L2 learners, such as French, German and Dutch learners [2, 3].

Many researchers have investigated this aspect of uncertainty (imprecision) and certainty (precision) [4, 5] by analysing texts produced by non-native speakers and

R.A. Thabet (✉)
Faculty of Education, The British University in Dubai, Dubai, UAE
e-mail: rawy.sabet@buid.ac.ae

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_32

contrasting them with native speakers' writing using certain software (viz., concordance software, such as WordSmith and Wmatrix3). The main approach used to hold this comparison between native and non-native features is called 'Contrastive Interlingual analysis' (CIA) [6]. This computer-aided method entails many functions, such as wordlist, which helps to find the words/phrases of high, medium frequency and even hapax legomena (i.e., words that are located only once in a corpus) [7].

In this current study, the researcher investigates the hedging markers used by the British University in Dubai (BUiD) students when they wrote their assignments for two modules (i.e., Research Methods in Education and TESOL Syllabus Design). 90 assignments, which formed the experimental corpus, were retrieved from Blackboard. All hedging markers and devices, which show the writer's uncertainty and certainty, were quantified and compared to another corpus written by native speakers who were at the same educational level. The two main questions that the study tries to answer are:

1. What are the most frequent hedging markers used by native and non-native writers?
2. Is there any significant difference between the frequencies of these markers in both writings?

Although there is an extensive literature on corpus analysis in other parts of the world, little research and investigation has been undertaken into postgraduate writing in the Arab Word, so this study seems to be one of the few sizeable corpora of tertiary English writing from the Middle East.

2 Literature Review

2.1 Introduction

First, the researcher discusses how the shift from accuracy to appropriacy has paved the way to the introduction of metadiscourse and corpus analysis. Then, the most relevant and seminal studies that discussed metadiscourse and how it is categorized are investigated. After that, hedges, as central exemplars of metadiscourse, are defined. Finally, the researcher explains how different researchers have approached hedges.

2.2 Metadiscourse

When there was a shift of focus from the mere study of language grammar to language function, metadiscourse found its way into this field of applied linguistics. The term metadiscourse was first introduced by Zelling Harris in 1959 (cited in 8)

to show how the writer guides the recipient to understand the text or speech in a certain way. The term has been used with other linguistic terms, such as connectives and hedges.

2.3 Metadiscourse Signals

Hyland [8] critically analysed the work done on metadiscourse and tried to present a more robust model, but ended up with a model that is very similar to Vande Kopple [9]. Hyland based his taxonomy on two main dimensions:

The interactive plane: On this plane, the writer is aware of the reader's anticipations and seeks hard to satisfy his/her needs and expectations using some resources (devices), which could be used to constrain/control what can be unfolded (understood) from the text by the reader [10]. This plane entails five categories: transition markers (e.g., and), frame markers (e.g., finally), evidentials (e.g., Z states, according to X), code glosses (e.g., such.) and endophoric markers (e.g., noted above.).

The interactional plane: On the interactional plane, the writer's stance and judgment can be clearly identified by the reader. The writer also creates an imagined dialogue with the reader and responds to the questions that the reader would raise. This plane entails five categories: hedges (e.g., about.), boosters (e.g., definitely), attitude markers (e.g., surprisingly), self-mentions (e.g., my), and engagement markers. It seems that the distinction between these two dimensions is vague and carries many interpretations. Both Hyland and Vande Kopple's models are very similar, but Hyland's model included more subcategories than Vande Kopple's (10 and 7, respectively). Additionally, it is more detailed and pays more attention to certain features such as how writers explain their stances and how they can engage readers. Hyland's list of hedges is of great importance to the researcher as he uses the same list and applies it to the two corpora. To be more precise, this list will be searched for in the two corpora to find how frequent each hedge is in the two corpora (native and non-native). Hyland's list of hedges consists of 101 hedges, but these devices were randomly mentioned on a list, so the researcher decided to improve this list by categorizing hedges according their part of speech (see Table 1).

2.4 Hedges

According to the Cambridge Dictionary [11], a hedge is "a word or phrase that makes what you say less strong". Hedging is a feature of academic writing which distinguishes it from other genres. There are some epistemic devices, such as *perhaps* and *may*, that show the open mindedness of the writer and that he/she does not have full commitment to what he/she is proposing. In other words, the presence

Table 1 Hyland's list of 101 hedges

Adverb		Adjectvie	Verb		Adjective phrase	Adverbial phrase	Modal auxiliaries	Noun
About = approximately	Plausibly	Apparent	Appear	Indicate	Certain amount	From my perspective	Could	Doubt
Almost = nearly	Possibly	Doubtful	Appeared	Indicated	Certain extent	From our perspective	Couldn't	
Apparently	Presumably	Plausible	Appears	Indicates	Certain level	From this perspective	May	
Approximately	Probably	Possible	Argue	Postulate		In general	Might	
Around = approximately	Quite	Presumable	Argued	Postulated		In most cases	Ought	
Broadly	Rather	Probable	Argues	Postulates		In most instances	Should	
Essentially	Relatively	Typical	Assume	Seems		In my opinion	Would	
Fairly	Roughly	Uncertain	Assumed	Suggest		In my view	Wouldn't	
Frequently	Sometimes	Unclear	Claim	Suggested		In this view		
Generally	Somewhat		Claimed	Suggests		In our opinion		
Largely	Typically		Claims	Suppose		In our view		
Likely	Uncertainly		Estimate	Supposed		On the whole		
Mainly	Unclearly		Estimated	Supposes		To my knowledge		
Maybe	Unlikely		Feel	Suspect				
Mostly	Usually		Feels	Suspects				
Often			Felt	Tend to				
Perhaps			Guess	Tends to				
				Tended to				

of hedges in writing proves that the information is subjective because it is given as a personal view rather than a fact [8]. Poos and Simpson [12] asserted that hedges can serve many pragmatic functions, for example, the hedging markers, such as *kind of* and *sort of*, can be used to show inexactitude (lack of precision) or to reduce the force of an ascertain. In a similar vein, Lakoff [13] describes hedges as those devices which make the writer's proposition fuzzier or less fuzzy. Learners of language should be taught how to strike a balance in their writing in order not to sound either arrogant or excessively tentative [14].

2.5 Meyer's Taxonomy

In a similar vein, Salagar-Meyer [15] proposed a taxonomy of hedging markers, which consists of five categories:

(1) **Shields**: this group consists of all auxiliary verbs (communicating possibility); lexical verbs with modal meaning such as *appear* and *seem*; adverbials of probability such as *likely*; adjectives of probability such as *probable*; epistemic verbs which are identified with the probability of a proposition such *to propose* or *to suggest*.
(2) **Approximators**: this category includes adverbs of degree, time and frequency, such as *approximately, roughly* and *often*. They are used to make things obscure or when the precise figures are inaccessible.
(3) **Author's personal point of view** (personal doubt), such as *I believe*.
(4) **Intensifiers** (emotional), such as *extremely interesting*.
(5) **Compound hedges**, such as *it could be suggested*. This subcategory can include compound hedges up to quadruple hedges or more, for example, *it would seem somewhat unlikely that*. Murniato [16] did a better job than Salagar as she (i.e. Murniato) divided the compound hedges into two categories: (1) a modal auxiliary with a lexical verb, which has a sense of hedging, for example *would appear* (2) a lexical verb with an adjective carrying a meaning of a hedge, for example *seem acceptable*. She added that these compound hedges could consist of double, treble or quadruple words.

In summary, all of these categories [Vande Kopple, Meyer, Lakoff and Hyland's) cause confusion and many of them overlap. For example, in Salager-Meyer's taxonomy and with a deep look at approximators and shields, it can be easily discovered that most of the approximators can do the same job of shields. In addition to that, many of the compound hedges consist of at least one main modal auxiliary, which is part of the shields. Koutsantoni [17] confirms that the examples of intensifiers given by Salager-Meyer are no more than examples of attitude markers and not hedges. She adds that the third category, which is the 'author's personal doubt', can include any item from the other four categories.

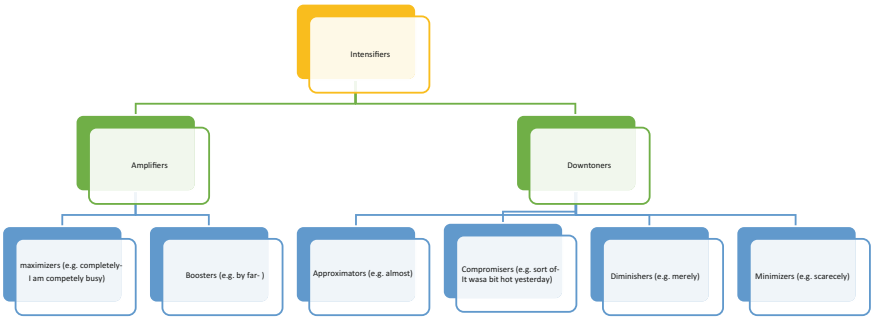

Fig. 1 Quirk et al.'s modal of intensifiers. Adopted from [18, p. 589]

Hedges can be expressed in different ways using different devices. Some of the devices that express the writer's engagement are boosters, diminishers and minimizers (adverbials of degree) (Fig. 1).

2.6 Intensifiers

Quirk et al. [18] distinguished between two main categories that show the writer's degree of commitment. These two categories are amplifiers (e.g., maximizers and boosters) and downtoners (e.g., approximators, compromisers, diminishers and minimizers-negative maximizers). Amplifiers are qualifiers or word intensifying expressions that reinforce the significance of adjacent expressions and show accentuation. Words that are usually used as intensifiers may include some adverbs, such as *completely* and *really*. If these intensifiers were ordered and distributed on an inverted triangle according the degree of emphasis, the maximisers sit at the top and the minimizers at the bottom (see Fig. 2). However, downtoners are words or expressions, which weaken the power of another word or expression. According to Quirk's categorization, the overstating is expressed by the amplifiers that show the

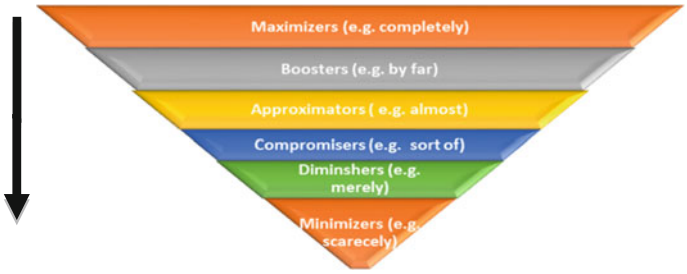

Fig. 2 Intensifiers pyramid

positive emphasis using emphatic devices while the downtoners are used to show the writer's caution. This caution is one the features that distinguishes the native speakers' writing.

2.7 Modality

Modality is usually connected with modal auxiliaries even though there are many other forms that would do the same function of modals, for example, modality could be expressed by some adverbs, such as *probably* and *possibly*; some verbs would also serve as modals, such as *I think* and *I feel* [19]. The previous studies, which were not based on corpus analysis, showed that non-native speakers tend to overuse or underuse certain modal auxiliaries/meanings [20]. Aijmer [19] used a computer-aided approach to compare between argumentative writings produced by Swedish L2 and English natives. What distinguishes her study is that she did not only compare between the Swedish L2's writing and native English speakers' writing, but she also held a comparison to the writing of other languages (i.e., French and German) regularly.

According to Papafragou [21] epistemic modality is defined as the "assessment of probability and predictability". Aijmer [19] used "degrees of likelihood" to explain the meaning of epistemic modality while root or deontic modality refers to the degree shown by the writer to express obligation, ability, power of deciding (volition), necessity, permission and necessity. Root modality has been referred to by many researchers, using different titles; for example, Halliday [22] refers to root modality as modulation.

2.8 Bundles

As mentioned earlier, these words, which usually occur together, have been given different titles such as clusters, bundles and multi-word expressions. This sequence of words helps us to identify the different registers, for example, *as can be discerned* refers to academic field and a bundle like *in pursuance of* refers to a legal document. The more proficient the writers become, the more bundles will be incorporated in their texts [23]. Wray [24] suggests that these formulaic patterns are overlooked in language acquisition. It is worth mentioning that these collocations can help to strengthen the relationship between the receiver of the text and sender because the presence of certain collocations helps the reader to know the register of the text.

2.9 Collocational Frame (It is … that)

Learners need to be equipped with the hedging devices that help them to strike "a balance between authority and concession" [12, p. 4]. As mentioned earlier in the literature review, in order for interlocutors to show their precision or imprecision, there are many approaches that they can use such as hedging markers or intensifiers (amplifiers and downtoners). What distinguishes an expert writer from an apprentice is the ability to vary the degree of precision to the extent that suites the context. Whether the interlocutor is hedging or boosting, his/her main objective is to comment on the proposition given by him/her. This comment could show how he/she feels towards what he/she is writing. This feeling could be related to the likelihood, the desirability or the seriousness of a proposition [25]. One of the evaluative forms that Lemke [25] studied was the sentences that include *It is…that*. Lemke [25] explained the use of *that* as a conjunction comes before a noun clause, whereas the extraposed *it is* precedes an adjective. This adjective could fall into one of seven semantic classes (probability, appropriateness, importance, seriousness, etc.). These adjectives are, in essence, evaluative epithets. He added that the noun clause that is introduced by *that* could represent a proposition or fact (if realis-) or a possibility (if irrealis). There is a variety of forms that this collocational frame could take, for example:

It + verb to be (functioning as a copula) + evaluative epithets (adjectives) + that…

Or *It + passive voice (to be + past participle) + that…*

These evaluative forms are very important for the study as the researcher examines them in both corpora and finally deducts some findings about their use, frequencies and varieties.

2.10 Lexical and Functional Words

There are two classes of words: lexical (also referred to as content or substantive words) and functional words. The former includes these words that carry meaning such as nouns, verbs, adjectives and prepositions [26]. They are also referred to as open-class category because it is possible for this category to be extended indefinitely by adding more items to it [27]. This idea can be supported by the fact that new words are coined and added to dictionaries almost every day. The second category includes the functional words, which is considered a closed-class list because there is a specific number of them and it is rare that new words are added to them. They serve as the mortar that sticks lexical words together [28]. When counting the elements that could be added to the functional word list, Hinojosa et al. [27] mentioned conjunctions, determiners, pronouns and prepositions. If the readers just go a few lines up, they will find that prepositions were counted among the lexical words by Corver and Van Riemsdijk [28]. This discrepancy in the

categorization corroborates the fact that the distinction between these two categories is not easy because some lexical words would serve as functional words and vice versa.

2.11 Corpus Linguistics Definition and Potential

Granger [6] states that corpus linguistics and second language research were two different fields, but with the advent of the new branch of knowledge known as learner corpus research in the 1980s, these two branches have been linked together. This new methodology has enabled researchers to explore different areas of language and make recommendations for better ways of learning a second language.

Corpus linguistics is defined as the analysis of electronic collections of authentic texts (i.e., naturally occurred). This authenticity feature was also mentioned by Halliday [29] as he enumerated three advantages and one disadvantage of corpus analysis. One of these advantages is that corpus enabled scholars to study grammar quantitatively. This quantitativeness is based on the ability of researchers to count the frequency of language items in texts [6].

Corpus linguistics is not a new method because it has been there for a long time, but with the advent of computers, this branch of study has enabled scholars to explore some areas that were very difficult to investigate without this magnificent device. The same idea of the added advantage of computers has been raised by Stubbs [30, p. 232] as he said, "the heuristic power of corpus methods is no longer in doubt". Although the focus of the corpus-based studies, conducted over the past two decades, was only on the features of the native English speaker, such as describing the registers and different dialects of Americans, British and Australians, this trend did not last for long as the focus had also been directed to non-native English. This change of focus started in the 1980s and the material collected from non-native English has been called learner corpora [6].

Halliday [29, p. 29] also defined corpus as "a large collection of instances of spoken and written texts". He added that the two main inventions that radically changed the work of grammarians are tape recorders and computers as the former was used to record the spoken discourse and the latter for saving the written texts. He continues to say that in the 1950s, when the two American scholars Randolph Quirk and W. Freeman Twaddell, started analysing their first corpus manually, they realized that the whole process would be computerized soon. Similarly, Schmitt [31] asserts that corpus analysis has recently gained significant popularity for two reasons: first, it focuses on the real language (spoken or written) produced by people; secondly, its outcomes can help in designing curricula.

3 Methodology

3.1 Participants

The participants of this study are students joining Master of Education programme, TESOL concentration. Each student has to study six modules (three elective and three core). The core modules are 'Teaching and Learning', 'Research Methods in Education' and 'Educational Policy'. The elective modules are 'Discourse for Language Teachers', 'TESOL Syllabus and Design', and 'Second Language Teaching and Learning'. The final written assignments, which were submitted to one of the core modules (i.e., Research Methods in Education) and to one elective module (i.e., TESOL Syllabus and Design), were uploaded to the corpus analysis software to be analysed. These two modules were carefully selected for the following reasons. First, the main question of this study is to find the frequency and quality of hedging devices used by the BUiD's students and comparing this frequency and quality to that of the BAWE writers. In 'Research Methods' module, students are required to write a research proposal while in 'TESOL Syllabus and Design' students are required to critically evaluate some syllabi. Therefore, in both modules students are expected to criticize the existing teaching material and methodology or to convince their study supervisor or funding institutions of the validity of their proposals. The total number of students that participated in this study is 70, who combined submitted 90 assignments. The number of assignments exceeds the number of students because some of them (20 students) submitted one assignment to each of the two modules. The majority of the participants are Arabs (85%) and 15% are from other nationalities, such as Indian (6%), British (2%), Bangladeshi, French, Nigerian and Pakistani with 1% each. The British participants were not raised in Britain, but were naturalized when they were adults. Finally, both genders were almost equally represented, as the male participants was accounted for 55% and female participants 45% of the study group. Ninety assignments were submitted to two modules—Research Methods in Education & Syllabus Design—between 3,000 and 4,000 words in length and with about 300,000 words in total. This number decreased to less than 300,000 when the text was formatted and converted to a text-only version, which is the appropriate format that can be uploaded to corpus analysis software.

3.2 Contrasting and Analysing the Two Corpora

According to Granger [6] contrastive interlingual analysis includes comparing NS to NNS. In this type, the contrast is held between writing features in both native (control) and non-native (experimental) English. The main concerns related to this type are the different varieties of native languages, such as the different dialects, spellings and the level of professionalism [32] of these native people whose writing

form the body of the control corpus. McKenny [14] stressed that for the two corpora to be successfully compared, the number of words and the purposes for which the texts of both corpora were written should match. This condition is met in this current study as both corpora have the same length and their texts are written to serve the same purpose. Contrasting native and non-native writing makes it possible to spot not only the misuse of some language features, but it also enables linguistics to determine the overuse and/or the underuse of some specific features when compared with native writing as a reference.

3.3 Motive Behind Writing and Corpus Compilation

McKenny [14] ascertains that most of the texts in the native language corpora were compiled for purposes other than corpus analysis. Similarly, BAWE corpus was made up of students' papers, submitted to their modules and not for corpus analysis. Generally, the subjects who contributed to BAWE and BUiD corpora were post-graduate students undertaking their master degrees. However, in BAWE case, students' papers were added to the corpus provided they gained a distinction. In addition, the authors of the selected papers were paid an amount of money and signed a disclaimer forms so that their universities could use their submitted paper for research purposes.

As for the compilation of the BUiD corpus, all word documents were converted to plain text because most tagging software works perfectly with texts that have no formatting (44). Since all section headings in the control corpus (BAWE) are encoded as < heading > … < /heading >, the researcher did the same thing in the experimental corpus. When the 90 assignments were joined using the WordSmith tool, each one of these assignments was given a specific number, for example, the first assignment was given the number 11, the second was given the number 22 and the last assignment was given the number 9090. Assigning numbers to each assignment would help the researcher to know in which assignment a specific language feature or concordance occurs.

3.4 Control Corpus Compilation

In order to obtain a full version of the British Academic Written English corpus, an online application form was completed and sent to the University of Oxford Text Archive. The request was soon approved and the researcher was given a full copy of the BAWE corpus. This corpus was compiled over a period of 3 years (2004–2007) and it consisted of 2,761 assignments written by students joining three universities; Oxford Brookes, Warwick and Reading [33]. All these writings were deemed as proficient writings (graded Merit or Distinction) and the authors were predominantly English native speakers (80%) and non-native English speakers (20%) [14].

The length of the texts ranged from 1,000 to 5,000 words. These written texts were classified into four disciplinary groups (DG), which are Arts and Humanities, Life sciences, Physical Sciences and Social Sciences. Then, the texts, submitted to each disciplinary group, were subcategorized into disciplines. Each disciplinary group consists of about 4–9 disciplines; for example, Arts and Humanities consists of 8 disciplines including Archaeology, Classics, etc. From all these contributions, the researcher selected texts submitted to Arts and Humanities and Social Sciences DGs. As the experimental corpus consists of assignments submitted to the Master of Education programme, the researcher tried to be very selective and had three main criteria when choosing the texts from BAWE. First, the topic had to be closely related to the educational field, such as English, History, Linguistics, and Sociology. Second, the more argumentative and text-oriented the piece of writing was, the more suitable it was deemed to be included for contrasting. Based on the previous criterion and based on the length of the experimental corpus (300,000), 101 texts were selected from BAWE with 300,000 words in total. All these key issues, such as the length and purpose of writing, should be considered when comparing the two corpora so that the only difference between them would be the level of proficiency and authorial expertise [34].

The focus of this study is the assignments written by 70 postgraduate students who undertook their master degrees at the British University in Dubai. The experimental corpus is referred to as the BUiD corpus. The methodology adopted in conducting this research is mainly empirical as it is based on direct observation of certain features in the two corpora (experimental and control). These features and language items have been quantified in the non-native corpus and then compared to the corpus written by native speakers. This method of contrast is called Contrastive Interlingual Analysis. As a starting point, all hedging markers, suggested by Hyland, were typed in a notepad to be searched for in both corpora. Homonyms, which do not serve as hedging markers, have been excluded. In other words, all language items that do not represent the writer's stance or degree of commitment are culled. In this regard, Aijmer [19] said that sometimes the manual analysis is necessary to avoid disambiguation. The manual filtering of both corpora, in this current study, resulted in deleting some markers that were mistakenly included within the list of hedges generated by WordSmith; for example, the epistemic meaning of the adverb *around* is approximately, but in concordance 1, it was used as a preposition which meant 'in this direction', so it was deleted. In concordance 3, the word 'May' served as the name of the fifth month of the year and not as a hedge, so it was deleted as well.

Concordance 1: *This gives more ..... of learning and competing* ***around*** *the world.*
Concordance 2: *There is …….* ***about*** *the tendency in …….*
Concordance 3: *April 2014–****May*** *2014 literature Review …….*
Concordance 4: *This reflected on ……; I felt helpless and defenseless.*

Annotating corpus is another solution to removing disambiguation, for example, tagging the word 'can' as a modal auxiliary when it serves as a modal and tagging it as a noun when it serves as a noun would help to distinguish between the auxiliary verb can and its homonym. To overcome the problem of unneeded language features, the researcher prepared a list of all search-words (hedging markers suggested by Hyland) and uploaded this list to WordSmith. A list of concordances of search-words was generated. The next step was filtering this list by deleting all irrelevant language markers or the markers that did not serve as hedging devices. Only the devices that showed tentativeness and degrees of un/certainty were included [35]. This step of weeding out devices that did not serve as hedging markers had been neglected by many studies as most of them followed "wanton frequency count" [12].

As mentioned in the literature review, Salagar-Meyer [15] did not develop a list of hedges for her proposed taxonomies, so the researcher referred to other studies to create a list for each taxonomy; for example, while reviewing the work of Hyland [8], it was found that the list of hedges entitled 'attitude markers', developed by Hyland, is very similar to the examples of intensifiers suggested by Salagar-Meyer. In the same vein, the researcher referred to the work of Holmes [36] to create a list of lexical verbs with epistemic meaning. Actually, this list was a merge of Holmes [36] and Hyland's (14) lists. Generally, most of these lists, used to search for concordances of Salagar-Meyer's taxonomies, were created in a similar way, i.e., merging the lists of hedges developed by other researchers to create one list for each taxonomy.

4 Findings and Discussion

4.1 The Most Frequent Single Words

As a starting point, the 40 most frequent single words were identified and compared in the two corpora (the experimental and control). The researcher started with single words and then moved on to compound forms. This sequence of steps is a representation of the bottom-up approach, which the researcher would like to follow in the beginning. According to Scott and Tribble [7] the most frequent words are found at the top while the tail of this list is full of hapax legomena. They also ascertain that once the text has been transformed into a wordlist, all the functional words, such as *the* and *of* are sent to the top of this list. As can be seen in Table 2, the first column contains the serial number of concordances; the second column shows the word itself; the third shows the number of tokens of each type of the words in the whole texts; and the extreme right-hand column shows the percentage of these tokens in texts as a whole. For instance, the word-type *the* has 22,979 tokens, which represents 7.55% of the whole running words in the BUiD Corpus. It can be easily discerned that there is a divergence in the use of the definite article *the* in both corpora: in the BUiD Corpus, the frequency of this article makes up 7.55% while in BAWE, it represents 6.88%. This finding contradicts McKenny's [14]

Table 2 The forty most frequent words in both corpora

N	BUiD			N	BAWE		
	Word	Freq.	%		Word	Freq.	%
1	THE	22,979	7.55	1	THE	20,781.00	6.88
2	AND	10,281	3.38	2	OF	12,543.00	4.15
3	OF	10,260	3.37	3	AND	9,104.00	3.01
4	TO	9,955	3.27	4	TO	8,271.00	2.74
5	IN	8,056	2.65	5	#	7,801.00	2.58
6	#	5,920	1.94	6	IN	7,042.00	2.33
7	A	5,486	1.80	7	A	6,192.00	2.05
8	IS	4,750	1.56	8	IS	5,010.00	1.66
9	THAT	4,055	1.33	9	THAT	3,511.00	1.16
10	STUDENTS	3,212	1.06	10	AS	3,226.00	1.07
11	BE	2,948	0.97	11	IT	2,321.00	0.77
12	FOR	2,823	0.93	12	FOR	2,182.00	0.72
13	THIS	2,685	0.88	13	BE	2,112.00	0.70
14	AS	2,616	0.86	14	THIS	2,062.00	0.68
15	ARE	2,430	0.80	15	WITH	1,881.00	0.62
16	IT	2,236	0.73	16	ARE	1,672.00	0.55
17	ON	2,113	0.69	17	BY	1,648.00	0.55
18	*WILL*	1,947	0.64	18	ON	1,643.00	0.54
19	WITH	1,929	0.63	19	NOT	1,553.00	0.51
20	TEACHERS	1,798	0.59	20	WHICH	1,521.00	0.50
21	THEIR	1,721	0.57	21	AN	1,384.00	0.46
22	THEY	1,429	0.47	22	FROM	1,276.00	0.42
23	BY	1,382	0.45	23	OR	1,185.00	0.39
24	LEARNING	1,382	0.45	24	WAS	1,087.00	0.36
25	LANGUAGE	1,358	0.45	25	*CAN*	993.00	0.33
26	STUDY	1,259	0.41	26	THEIR	968.00	0.32
27	RESEARCH	1,188	0.39	27	HAVE	907.00	0.30
28	WHICH	1,123	0.37	28	I	891.00	0.29
29	NOT	1,115	0.37	29	HIS	852.00	0.28
30	TEACHING	1,078	0.35	30	BUT	834.00	0.28
31	HAVE	1,071	0.35	31	HAS	827.00	0.27
32	AN	1,063	0.35	32	P	813.00	0.27
33	OR	1,054	0.35	33	AT	800.00	0.26
34	FROM	1,010	0.33	34	MORE	781.00	0.26
35	*CAN*	990	0.33	35	THEY	781.00	0.26
36	LEARNERS	933	0.31	36	ONE	768.00	0.25
37	TEXTBOOK	906	0.30	37	HE	698.00	0.23
38	SCHOOL	872	0.29	38	ITS	660.00	0.22
39	BOOK	866	0.28	39	*WILL*	655.00	0.22
40	TEACHER	846	0.28	40	ALSO	643.00	0.21

conclusions as he reports that the non-native speakers in his study significantly underused the definite article when compared to the native speakers. The definite article usually collocates with nouns. To prove that, when the definite article is searched for in the BUiD corpus, it is found that it collocates with the word *STUDENTS* 871 times. This finding suggests that there would be an overuse of nouns in NNSs' corpus. This will prove right when the two corpora are tagged with the USAS tagset. As an ESL teacher with many years of experience teaching Arabs, the researcher can assume that the overuse of the definite article is due to its wrong use, which could be attributed to the L1 transfer.

As is expected, the most frequent words on the top of both lists are functional words such as *the*, *and*, *of* and *to*. It is worth mentioning that the top 9 most frequent words are almost the same in the two corpora. It is also interesting to notice that on the experimental list (BUiD Corpus), the first content word comes in the tenth position while there is no one content word among the 40 most frequent words in the reference corpus as all of these 40 most frequent words are functional words. It is equally interesting to notice that the frequency of the modal verb *will* is 1,947 while it is only 655 in the BAWE corpus.

4.2 Lexical Density

Lexical density is usually used as a measure of the level of proficiency of text. Kenny [37] developed a technique that is referred to as the type-token ratio (TTR). As the name indicates, the total number of word types is divided by the total number of the running words. Then, the result of this division (i.e. quotient) is converted to a percentage. This technique had a lot of criticism because of its sensitiveness to the length of the text, for example, if a text consists of 10,000 running words, it is said that this text has 10,000 tokens. This dependence on the size of the text is considered one of the limitations of this measure, which could have been firmly accepted if it had excluded the repeated words (Table 3).

In his endeavour to overcome the deficiency of the TTR, Scott [38] developed the standardized type/token ratio by dividing the texts into smaller segments and taking the average of the TTR of each of the segments. This approach was also criticized for not reflecting the reality of the text lexical density.

In reaction to the limitation of both techniques (i.e., TTR & STTR), scholars started to adopt another tool developed by Ure [39]. In order to measure the density of lexis in a text, Ure tried to find the proportion of the lexical words to the grammatical ones. As recommended by Ure [39] and Stubbs [30], the lexical density is calculated by dividing the lexical/content words by the total number of

Table 3 TTR & STTR of the two corpora

Corpus	BUiD	BAWE
Tokens (running words) in text	304409	302121
Tokens used for word list	298489	294320
Types	10904	17607
Type/token ratio (TTR)	3.65	5.98
Standardized type/token ratio (STTR)	37.53	40.54

tokens in the corpus. To create a list of content words, the researcher used a stoplist of the 100 most frequent words.

In the BUiD Corpus (using the 100 most frequent words)

Content words = 304,409 (tokens) − 143.445 (functional words removed) = 160.964….

Lexical density = 160.964 (content words)/304.409 (tokens) = 52.877%.

In the BAWE Corpus (using the 100 most frequent words)…Content words = 302.121 − 141.683 = 160.438.

Lexical density = 160.438/302.121 = 53.103%.

The percentages in the Table 4 suggest that the lexical density of the native speakers' corpus is slightly higher than the non-native's. This is not a surprising finding for the researcher because he expected that the lexical density of BUiD would be less than BAWE, because he is an Arab and was educated in an Arabic country where teaching is mainly grammar-oriented. However, the high lexical density is not evidence of the full command of the language as there are native speakers whose writing is not highly lexically dense [40]. In addition, the categorization of a text into lexical and functional items is not easy because some lexical words work as grammatical words and vice versa [41]. In other words, the function of each category (lexical and grammatical) may overlap.

4.3 Hyland's Taxonomy

As mentioned earlier in the methodology section, all hedging markers (101), suggested by Hyland [8], were typed in a notepad and searched for in both corpora using the function of 'get search-word from a file' in the corpus analysis software called 'WordSmith'. When adding all totals of hedging adverbs, verbs, adjectives,

Table 4 Lexical density using a stoplist of the 100 most frequent words

Corpus	BUiD (%)	BAWE (%)
Lexical density using stoplist	52.877	53.103

Table 5 Hedges according to Hyland taxonomy

Part of speech	BUiD	BAWE
Hedging modal auxiliary	1661	1617
Hedging adverbs	723	1096
Hedging verbs	691	1026
Hedging adjectives	94	200
Adverbial phrase	66	50
Hedging noun	11	20
Noun phrase	5	13
Total	3240 (1.07%)	4002 (1.33%)

Table 6 Expected contingency

Part of speech	BUID	BAWE
Hedging modal auxiliary	1460.29	1803.73
Hedging adverbs	810.334	1000.91
Hedging verbs	764.895	944.787
Hedging adjectives	130.972	161.775
Adverbial phrase	51.6761	63.8295
Hedging noun	13.81	17.0579
Noun phrase	8.0187	9.90458

modal auxiliary and compound hedges, it was found that, generally, the NSs used more hedges than NNS; 4,022 (1.33%) hedges and 3,251 (1.07%) hedges, respectively (Tables 5, 6).

Chi-square = 3.14… Degree of freedom (df) = (C−1) (r−1) = (2−1) (7 −1) = (1) (6) = 6 Probability = 0.05.

Based on the Chi-square results, there is a likelihood that there would be a statistically significant difference between the frequencies of the hedging markers in the two corpora. Looking closely at the frequencies of hedging markers in both corpora, it can easily be discerned that NSs employed more adverbs of probability than NNSs, especially, the adverb 'perhaps' which was used 81 times by NSs while the NNSs used it only eight times. Similarly, adverbs like 'possibly', 'likely', 'roughly' were far underused by the NNS.

4.4 Salagar-Meyer's Taxonomy

When applying Salagar-Meyer's [15] proposed taxonomy of hedging markers, which consists of five categories, it was also found that native speakers, overall, used more hedging devices than non-native speakers; 9,945 (3.37% of the total number of words) and 7,324 (2.42%), respectively (see Table 7). The two most frequent types of hedges in both native and non-native speakers' corpora are shields and author's personal point of view. These two types accounted for 52.81 and 36.65% of the total number of hedges used by native speakers whereas they

Table 7 Salagar-Meyer's proposed taxonomy of hedging markers found in BAWE and BUiD

Category	BAWE				BUiD			
	Freq	Percentage of this category to the total # of token	Percentage of this category to the total # of hedges	Expected contigency	Freq	Percentage of this category to the total # of token	Percentage of this category to the total # of hedges	Expected contigency
Shields	5252	1.8	52.81	5373.0297	4078	1.35	55.68	3956.9703
Approximators:	381	0.13	3.83	393.907	303	0.1	4.14	290.093
Author's personal point of view	3645	1.22	36.65	3268.737	2031	0.67	27.73	2407.263
Intensifies (Similar to attitude markers developed by Hyland 2005)	632	0.21	6.35	884.56309	904	0.3	12.34	651.43691
Compound hedges such as it could be suggested.	35	0.01	0.35	24.763159	8	0	0.11	18.236841
Total	9945	3.37			7324	2.42		

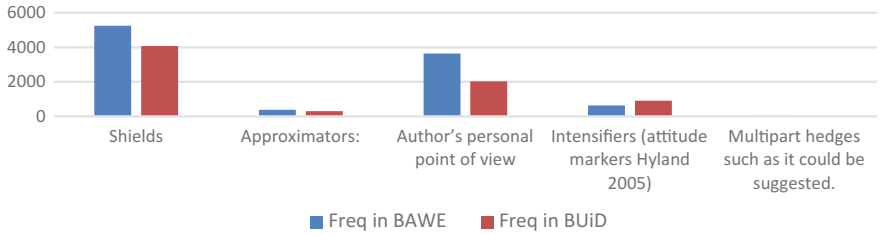

Fig. 3 Hedges in BAWE & BUiD (using Salagar-Meyer's taxonomy)

constituted 55.68 and 27.73% of the total number of hedges used by non-native speakers [42]. The native speakers exceeded the non-native speakers in the frequency of the shields, approximators, author's personal point of view and compound hedges. The order of the hedge types in both corpora is the same as shields come in the first place and Author's personal point of view come in the second place followed by intensifiers, Approximators and compound hedges (Fig. 3).

Chi-square = 1.9 > 0.05 Degree of freedom (df) = (C−1) (r−1) = (2−1) (5−1) = (1) (4) = 4 Probability = 0.05.

Based on Chi-square result, there is a significant difference between NS and NNS in their use of hedges. Generally, this taxonomy (i.e. Salagar) is problematic, especially, the category of intensifiers which was described by Koutsantoni [17] as vague and function as attitude markers more than as hedges. Based on this conclusion, the researcher used Hyland's list of attitude markers as intensifiers. This vagueness and lack of a list of lexical items led to a discrepancy in the counts of intensifiers calculated when Salager-Meyer [15] and Quirk et al.'s [18] models were applied.

4.5 Syntactic and Semantic Tagging

The researcher also used Wmatrix3 to identify the variety of parts of speech used in both corpora. The two corpora were uploaded to the Wmatrix3 tool and tagged with the UCREL CLAWS7 tagset. The main motive behind this step was to find whether the non-native speakers in the experimental corpus overused or underused some parts of speech. It is clear that NNSs used more verbs, nouns and fewer adverbs and adjectives than NSs (see Table 8). This finding is almost in line with Ringbom [43] who found that NNSs corpus included more verbs and fewer adjectives than NSs'.

Table 8 Parts of speech in both corpora

Total number	BUiD	%	BAWE	%
Verbs	53713	18.64	46476	16.14
Noun	84837	29.41	80320	27.89
Adverb	9998	3.42	13788	4.77
Adjective	23094	8	27555	9.56
Total	**171642**	**59.47**	**168139**	**58.36**

Table 9 Parts of speech-CLAWS tagset

Sr no	Item	O1 (BUiD)	%	O2 (BAWE)	%		LL	Logratio
1	NN2	27735	9.62	16877	5.86	+	2667.5	64.31
13	**VM**	**4748**	**1.65**	**3533**	**1.23**	**+**	**178.7**	**34.37**

The researcher was looking for the devices and parts of speech that were used to show tentativeness or degree of commitment. According to the CLAWS7 tagset, VM stands for modal auxiliary (e.g., can) and that was the first target for the researcher. Searching the tagged lists of the two corpora, where O1 stands for observed frequency in the BUiD corpus and O2 stands for observed frequency in the BAWE corpus, VM was the thirteenth item on the list (see Table 9).

Modal auxiliaries have significant importance, as their proper use by non-natives is considered to be a challenge. They are also important devices used by the writer to show tentativeness or hedging [44]. For these two main reasons, the researcher decided to investigate modal auxiliaries in both corpora. It is clear from Table 9 that BUiD students overused the modal auxiliaries as there are 4,748 occurrences of them, which represent 1.65% of the total words in BUiD while BAWE students used them 3,533 times, which represent 1.23% of the total words in BAWE. These numbers are different from Hyland's because Hyland's list of modal auxiliaries did not include can and will. This finding (i.e., the overuse of modal auxiliaries by non-native speakers) necessitates having a deeper look at the different modal auxiliaries and investigating them individually to find the reasons behind this tendency (Table 10).

Examining the frequency of the modals, it was found that *will, can, should* and *might* were overused by NNSs while would, *may, could, must, shall*, *can't* and the

Table 10 Frequency of modal auxiliaries

Word	Frequency in BUiD corpus	Relative frequency	Frequency in BAWE corpus	Relative frequency
will	1921**	0.67	620	0.22
can	1035	0.36	1097	0.38
should	530*	0.18	313	0.11
would	374	0.13	486	0.17
may	273	0.09	338	0.12
could	255	0.09	319	0.11
might	191*	0.07	125	0.04
must	137	0.05	192	0.07
shall	23	0.01	20	0.01
can't	4	0	11	0
need	2	0	10	0
can not stand	2	0	2	0
Shall/will	1	0	0	0
Total	**4748**	**1.65**	**3533**	**1.23**

semi-modal *need* were underused. In the BUiD corpus, the modal auxiliary *will* came in the first place with the highest number of frequencies (1,921 times) followed by *can* with 1,041 occurrences while in the BAWE corpus, the order is reversed where *can* occupied the first place with 1,097 occurrences and will the second place with 620 occurrences. Generally, within the global list of BUiD, it is the modal verb *will* that mostly stands out because BUiD students used this modal almost three times as often as BAWE students. This overuse could be attributed to either L1 transfer (interlingual), developmental factors, or speech-like writing (viz, students' writing is affected by the way they speak, i.e., register-interference). The last reason needs to be supported by referring to an Arabic corpus where this feature of modality can be checked. Another reason, I would suggest, could be that in one of the modules, Research Methods, students were requested to write a research proposal, and so they used the word *will* many times to talk about their plans even though the present simple could have been used to express future planned activities. For example, one of the students was discussing the approvals that he would get to be able to run his study said, "*an approval on the study will be obtained from the HCT research*"; *someone else who was explaining the stages of his research said "[t]he first stage will involve questionnaires to be collected*". A final potential reason for the overuse of *will* by NNSs is that it is teaching-induced. It has also been noticed that NSs used *may, could* and *would* (modals mainly express probability) more than NNSs. This could be attributed to the fact that NSs tend to use these modals when they wish to show their attenuation about their propositions (epistemic stance). Finally, the modal auxiliary *should* is one of the modals that was overused by BUiD students. When the researcher examined the occurrences of *should*, he found that students used this modal mainly to express the ethical code of conduct or norms that usually prevail in teaching and research contexts; for example, "*the teacher should aim to create a suitable psychological atmosphere in order to lower learners' anxiety arising from their increased autonomous roles*".

4.6 Modals with Deontic and Epistemic Meanings

The next step for the researcher is to find out how many of these modal verbs have deontic meaning and how many have epistemic meaning. As mentioned earlier, the researcher follows the bottom-up approach when analysing the devices used to express modality. In other words, he starts with analysing the single modal auxiliaries, then the modal adverbials and finally the harmonic modal combinations. Before exploring the different modal auxiliaries, used by both NSs and NNSs, it is important to discuss the root and epistemic meanings of modal verbs. Although there seems to be a unanimous agreement among researchers on the forms of modal verbs that are used to express modality, Coates [44] and Hermeren [45] compiled a list of modals other than the one agreed upon by most researchers. They adopted a different technique, which is based on ferreting out the frequency of the various modals. This approach required putting a lot of effort and time because they had to

check every single form to find whether it serves the epistemic or root meaning [36]. Each one of the two corpora (used in this current study) consists of about 300,000 words and with this big size of the corpora, the researcher decided to investigate the function (epistemic or root) of only some of the modal. It was also very helpful to refer to Aijmer's [19] study in which she classified modal auxiliaries as follows: *Must, may, should* and *might* could have both root or epistemic meanings while *have to, must, ought to* and *should* usually serve as root modals; the remaining verbs like *will, would* and *could* usually serve as epistemic devices; the modal verb *will*, in particular, is used to express the future plans, but with some kind of certainty. Although the number of modal auxiliaries is few, it seems to be a challenging task to determine the function of these modals because they are polysemous, for example, *can* is used to express possibility, ability and permission [18]. In this study, the researcher intends to investigate only the epidemic and root meaning of the modal auxiliary *would*, which would somehow show the preference of native and non-native towards the use of epidemic and root meanings of modal auxiliaries.

Table 11 shows all the occurrences of *would*. The epistemic modal *would* followed by *be* was underused by BUiD students as they used it eight times less than BAWE. The combination of *would* and verb *to be* was usually followed by an adverb (*would be very useful*), or past participle (*would be given*) or present participle (*would be asking*) or an adjective (*would be ideal*) or prepositional phrase (*would be of great help*). *Would* was mainly deployed in both corpora as the epistemic modal except for some forms, such as *would like* which served as a polite way to request something. BUiD students significantly overused this form, which carries the root meaning of would. This finding corroborates the previously proven fact, which suggests that BUiD students tend to hedge less than their counterparts in BAWE. Additionally, the modal verb *would* can be used to express probability or the possibility per se, not to mention adding another lexical verb with epistemic meaning like *appear* or *seem*. This combination strengthens the meaning and shows that the writer is trying to be objective as much as possible. Examining this combination of *would* and some lexical verbs with epistemic meaning like *seem, appear*, and *need,* it can be easily discerned that BUiD students significantly underused this combination of double hedging.

Table 11 'Would' with root & epistemic meanings

Item Would	BUiD	BAWE	Root or epistemic
Would be	135	143	Epistemic
Would better	2	2	Epistemic
Would like	26	11	Root
Would + adverb	28	58	Epistemic
Would + seem/appear/need	3	25	Epistemic

4.7 Harmonic and Disharmonic Combinations of Modals and Adverbs

Sometimes the modal verbs interplay with other lexical verbs or other parts of speech, which perform the same function of modal auxiliaries [46, 49]. For example, the will certainly combination of the modal and adverb is considered harmonic because the modal auxiliary will is used to denote certainty in the future and the adverb certainly strengthens the certainty of the verb will. Examining Table 12, it can be easily discerned that both NNSs and NSs used the harmonic modals *would probably* and *would definitely* equally, but the NSs used *would surely* four times more than the NNSs. Similarly, the combination of *will likely* was used by NSs twice as much as NNS, but the combination of *will most likely* was not seen in the NSs' corpus. Generally, NSs used more combinations than NNSs. Contrary to this finding, Aijmer [19] concluded that NNSs used more combinations and with different types and she attributed that to either the influence of spoken language or the L1 transfer.

4.8 Intensifiers

In this current study, the researcher could identify these intensifiers (adverbials of degree) using the Semantic Tag function and USAS (UCLER Semantic Analysis System) on Wmatrix3. This tool helps to group word senses together and categorize them according to the generality they lie within [47]. According to USAS tagging, each word within the two corpora is assigned a semantic and syntactic tag. This approach makes it easy to identify the behaviour of words like adverbials of degree. When the semantic tags of the two corpora were juxtaposed, it was found that NNSs underused all of the adverbials of degree except for the approximators. It did not seem wise to conclude that the low count of adverbials of degree implies that non-native speakers' writing was less proficient. In other words, it was too early to judge that low/high frequency stood for low/high proficiency in writing, but it was worth having a deeper look at the different patterns of adverbials used by both NSs and NNSs and trying to justify their under- or over-use. As can be seen in Table 13, there is a statistically significant difference in the count of adverbials between NSs and NNSs. The former used some adverbials almost twice as often as the latter, but both NSs and NNSs used maximizers almost equally as there is no significant

Table 12 Harmonic and disharmonic of modal interplay

Modals interplay	BUiD	BAWE
Would		
Would probably	2	2
Would surely	1	4
Would definitely	2	[2]

Table 13 Adverbials of degree (USAS)

	Amplifiers		Downtoners			
Corpus	A13.2	A13.3	A13.4	A13.5	A13.6	A13.7
	Maximizers	Boosters	Approximators	Compromisers	Diminishers	Minimizers
BAWE	476	1496	189	129**	262**	122**
BUiD	461	1076	195	57	109	53

difference between them with the log-likelihood (LL) = 0.24 which is less than the LL cut-off at 6.63. However, the difference between the frequency of boosters is significant as the LL = 68.97 which is higher than the cut-off value. NNSs are often stigmatized for their overstatement and use of boosters, but in this case, it proves the opposite as the NNSs underused almost all the scalar intensifiers [18].

This finding (i.e., underuse of amplifiers) is congruent with Granger's [48]. In order to find the reason behind this underuse, she added up the total number of tokens of amplifiers (including both maximizers and boosters) and the total number of types of these amplifiers. To her astonishment, she found that NNSs underused both the types and tokens of amplifiers. The low number of types could be expected, as the NNSs, unlike the NSs, do not have a rich language variety at their disposal. However, the second finding, which is the low number of tokens, is surprising as this means that NNSs' language is less emphatic or hyperbolic than NSs. This last conclusion contradicts the well-known thought, which implies that NNSs tend to overstate issues more than NSs [32]. As mentioned earlier, the findings of this current research, pertaining the tokens and types of the amplifiers (see Table 14) found in both corpora, are consistent with Granger's. Therefore, the researcher decided to investigate the frequencies of boosters in the two corpora to find out which boosters the NNSs underused or which ones they did not use at all. Boosters, in particular, were focused on and investigated in detail because they were the main reason of the high frequency of amplifiers in both corpora. When the lists of boosters were compared, it was found that there are 22 types of boosters (with 38 frequencies in NSs' corpus) that were not deployed at all by the non-native speakers (e.g., remarkably, desperately and agonizingly). As mentioned before, this case of non existence of some boosters in the NNSs' corpus could be attributed to the "natural deficiency of non-native vocabulary" [32, p. 28]. Similarly, most of the boosters, underused by non-native speakers, were a combination of an intensifying adverb ending with the suffix -ly followed by an adjective (adv-adj-, e.g., increasingly difficult). This type of adverbial collocations requires high combinatory skill, which is not within the capabilities of the non-native speakers. To counteract this deficiency, NNSs resorted to use all-round/stereotyped boosters that

Table 14 Types and tokens of amplifiers

Amplifiers	Types		Tokens	
	NS	NNS	NS	NNS
Maximizers	32	24^{-}	476	461^{-}
Boosters	50	34^{-}	1496	1076^{-}
Total	82	58^{-}	1972	1537^{-}

can be used in many contexts, such as very. This booster was overused by NNSs as they used it 272 times while NSs used it 187 times only.

Looking closely at the occurrences of some other boosters, it is found that NSs used more complex forms of some boosters than NNSs; for example, more, the most frequent booster in both corpora with 531 occurrences in BAWE and 441 in BUiD, was used in compound forms with a sense of a downtoner, such as "which was no more than a form of collective identity" and "the world today is no more than a global triumph of free market". However, when the researcher examined all the occurrences of more in BUiD's corpus (NNSs), no one example of such a complex form was found. Most of, if not all, cases in which more was used, were comparisons, such as "the findings will be more reliable" and "to write more details". This means that NSs have the linguistic competence that enables them to use words in more varied and complex forms than that of the non-native.

In addition to that, in the NSs corpus, with close investigation of the occurrences and contexts in which more was used, it was found that most of the cases denoted understating more than overstating. In other words, Wmatrix3 misinterpreted these devices as boosters, but in reality, they were no more than expressions of understatement.

As for the frequencies of diminishers and minimizers, NSs far exceed the NNSs in the use of these downtoners. This means that NSs were more cautious than NNSs as the former used the downtoners devices to show some kind of vagueness, which is now considered one of the main characteristics of the native speakers' language [49]. However, the NNSs used more approximators (195) than NSs (189) (see Table 14). Although the difference was not great, it proved that NNSs sounded more tentative than NS.

It is also worth mentioning that NSs' use of compound downtoners far exceeded the NNSs, for example the diminisher to some extent was used by the NSs twice as much as the NNSs (11 times and 4 times, respectively). This corroborates the fact that NSs have the ability to form varied and complicated structures of language items, even the hedged ones.

4.9 State of Inexactitude

According to Quirk et al. [18], *sort of* and *kind of* are considered part of the compromisers, but they were not included in the list generated by Wmatrix3 (USAS function), so the researcher decided to search for them using Wordsmith and the results are shown below.

As can be seen in Tables 15 and 16, there are 17 concordances of sort of in the BAWE corpus and seven concordances in the BUiD corpus. Some concordances of *sort of* in the two corpora were culled in order to exclude all the examples, which did not serve as a hedging marker. For instance, in Table 16, line number 4 was deleted because the phrase *sort of* in this context was a synonym of type of and it did not have the sense of a hedging device [12]. It is worth mentioning that while

Table 15 Concordances of sort of in BUiD corpus

Concordances of sort of in BUiD corpus
1. with the receptive skills as a **sort of** warming up for the productive skills
2. assume that there should be a **sort of** reconsideration of the number of
3. tha~~t's implemented directly from~~ **sort of** ~~answers which will determine~~
4. learning L2 and establish some **sort of** a bridge between both language
5~~. vidual on the planet has some~~ **sort of** ~~a gadget that connects him/her to~~
6. n of the book therefore, such **sort of** question helped in establishing the
7. ~~establishing ICTs within this~~**sort of** ~~perform rather than other people~~

Table 16 Concordances of *sort of* in the BAWE corpus

Concordances of sort of in BAWE corpus
1. inspiration". The writer takes on a **sort** of god-like essence as Author
2. of literary production as "a **sort of** involuntary secretion" described by
3. stitutional change - causes a **sort of** national reappraisal of institutions
4. ~~, such as nails, ironworks, a~~ **sort of** ~~mortar and some kind of candles.~~
5. to justifiably attribute any **sort of** idealism to Husserl, the evidence is
6. ~~Scope ambiguityThis is the final~~**sort of** ~~ambiguity which is caused by~~
7. the very heart by a pleasant **sort of** involuntary helplessness" and yet "
8~~. things." Correlatively, the same~~ **sort of** ~~optimism is just as comical~~
9. ~~had been used to uphold some~~ **sort of** ~~roof of which just a few pieces~~
10. the way it is because of some **sort of** intending or pointing on behalf of
11. posed that "Children use some **sort of** nonsemantic procedure to
12. ~~Nietzsche an intentional choice, the~~ **sort of** ~~absolute undecidability~~
~~13. offer prior justification for the~~ **sort of** ~~cognition that can come to know~~
~~14. guage barrier, is exactly the~~ **sort of** ~~reality people with hearing~~
~~15. with impairments (Oliver, 1990). The~~ **sort of** ~~approach which is evident~~
16. . The inference was that this **sort of** 'being inside something and looking
~~17. English (Roach 2000). Thus this~~ **sort of** ~~group is called tone unit which~~

the researcher was weeding out the examples of sort of in the BAWE corpus, which did not have the sense of a hedge, he did not find it a challenging task. However, when he carried out the same task in the BUiD corpus, it took him more time to distinguish between the examples of both types and meanings of sort of, which could induce a kind of unsuitability of the use of these hedges. After weeding out the non-hedging examples of sort of, it was found that NSs used this hedge twice as much as NNSs. This finding gives another evidence that NSs tend to show their

tentativeness by using these expressions of inexactitude that would invite the reader to take part in the debate being initiated by the writer. In other words, the writer tries to play the role of the reader by judging his/her own stance and determining to what extent he/she (i.e., the writer) is true or false.

4.10 Collocational Frame (It Is … that)

The two forms below and any other form that represented the writer's stance were searched for in the two corpora on WordSmith, using the collocational frame *it ** that.*

It + verb to be (functioning as a copula) + evaluative epithets (adjectives) + that… Or…. *It + passive voice (to be + past participle) + that…*

Then all the concordances that did not represent the writer's stance, were culled using 'Delete' and 'zap' functions in the WordSmith tool. Here are some examples of the culled concordances below. The first example (Concordance 6) was mistakenly included because the tool did not distinguish between the extraposition it is…that and any other form that included the adjacent words 'it…that'; this was the reason for including the first example in the concordances on WordSmith. The second example (Concordance 7) suggests that this student was not aware of the different correct forms of the extraposition and this explains the reason for entering incorrectly the adverb 'clearly' in place of the adjective 'clear', which should have been used here. The other examples contain the pronoun it, which functioned as an object for a verb and not as a part of the extraposition collocational frame. Additionally, concordance number 7 represents a case of it-clefted (Table 17).

Concordance 6: *supported it with diagrams that*
Concordance 7: *It is clearly that through this method*
Concordance 8: *Define **it** as"…a process that*
Concordance 9: *Merely choosing a textbook without first evaluating **it** would mean that*
Concordance 7: ***It** is there that he writes*

NNSs used more extraposed collocational frames than NSs, with usages of 124 and 106, respectively (142 and 125 concordances before culling). However, the quality, variety and complexity of the structures that come after the expletive it in NSs' concordances, are more advanced than NNSs and show how competent the native speakers are. Some of the most advanced expressions used by the NSs are it is poignant that; it is ironic that and it is posited that. None of these adjectives (i.e., ironic and posited) were used by the NNSs. As can be seen in the tables above, the concordances were sorted by the percentage of frequency of each one of these extrapositions. In BUiD, the extraposition *It is found that* comes in the first place with 99% of the whole texts while the extraposition that occupied the first place in BAWE, is *it is likely that*. The modal adverb *likely* was defined by Salagar-Meyer

Table 17 The first 20 concordances of the extraposition 'it**that' in BUiD and BAWE

N	Concordance in BUiD	Word #	%	N	Concordance in BAWE	Word #	%
1	interaction in the classroom. **It is found that** there are actually several weak	4168	99.00	1	Generative Grammar framework, **it is likely that** the minor differences of perspective	1924	99.00
2	through the different tests. So **it is recommended that** this contradiction	4012	99.00	2	sensitise educators; however **it is doubtful that** students need to be aware of	3304	99.00
3	appropriate for them Secondly, **it is crucial that** the authors would use more	4090	98.00	3	s from other genres. However, **it is likely that** most texts will still aim to be	4029	99.00
4	to make this book more useful **it is recommended that:** 1) A needs analysis	4160	98.00	4	ace'. Thus, with this in mind **it is hoped that** with time some inroads may be made	5345	99.00
5	forts to reach it Generally, **it is thought that** adhering to the supplies of	2392	97.00	5	and interrogative sentences. **It is certain that** this area will present the logician	3225	99.00
6	impressive and meaningful. **It is said that** practice makes a man perfect	3945	97.00	6	tak, 1990: 351). In addition, **it is poignant that** Nisa herself chose the name	4981	98.00
7	in the textbook. Furthermore, **it was found that** the dominance of the listening	4227	96.00	7	very nature of its structure, **it seems unlikefy** that English will be ousted in favour of	1877	98.00
8	unspecified forms in instruction. **It was argued that** such way will cause	4010	96.00	8	biggest ever budget in 1944. **It is certain, that** before the war had ended	27 60	96.00
9	. 5.Conclusion To conclude, **it is clear that** whatever is called a paradigm	3878	93.00	9	the world. However, although **it is true that** Musil's descriptions of the Other	5363	96.00
10	reading, and writing. However, **it is hypnotized that** teachers employ the	3491	93.00	10	. From reading Shostak's text **it becomes apparent that** it was as much about her	4819	95.00
11	, rank it as totally lacking. **It is noticeable that** the	3472	93.00	11	ing styles in modern theatre. **It rings true**	3924	95.00

(continued)

Table 17 (continued)

N	Concordance in BUiD	Word #	%	N	Concordance in BAWE	Word #	%
	textbook does not allocate				**that** action is louder than words		
12	rom the result of this study, **it is concluded that** integrating such aids with	4155	92.00	12	s intellectual^ bankrupt and **it is claimed that** social identities are created by	2820	93.00
13	in a sentence. For all above, **it is concluded that** the UAE English skills textbook	3836	91.00	13	lly promoted to children, but **it was discovered that** it it appealed to both children and	3224	91.00
14	ve their progression. Likewise**,it was perceived that** using of blogs helps	3094	91.00	14	ted to insincere conclusions. **It is possible that** Bull weighted his analysis in favour of	5954	91.00
15	appendices C & D). Finally, **it was noticed that** the units'themes are of little	3936	90.00	15	is the "hypothesis testing". **It is assumed that** output provides learners with the opportunity	3670	90.00
16	to the cultural restrictions. **It is recommended that** this study can be carried	2899	90.00	16	ernet transactions." (URL). **It is ironic that** most of the content available on the Internet	1709	89.00
17	otions in effective teaching **It is argued that** assessment guidelines and	3356	90.00	17	qualsiasi are stressed), and **it would seem that** if these linguistic alternatives continue	2822	87.00
18	adictoiy to this approach. So **it is considered that** such an an experiment	3777	87.00	18	less, as Lyons (1977) argues, **it is clear that** there are strong semantic associations	1568	87.00
19	listening to writing Thus, **it is important that** teachers introduce lessons	3422	87.00	19	th Tyson's 'architect' model. **It was recognised that** the need for the roles of 'clerk of works'	2926	87.00
20	ned the problem faced in UAE. **It is evidenced that** most of the students	3624	86.00	20	oncrete groups as they stand. **It is clear that** whichever scenario is true, the Theban Magical	4208	86.00

Table 18 Examples of adjectives of importance and probability

Adjectives of	BUiD Corpus		BAWE Corpus	
	Examples	Frequency	Examples	Frequency
Importance	It is important that	4	It is important that	4
Total 1		4		4
Probability			It seems likely that	2
			it seems unlikely that	1
			It is possible that	4
			It seems possible that	1
			it is likely that	2
			it is unlikely that	1
			it is doubtful that	1
Total 2		0		12
Total of total		**4**		**16**

[15] as one of the shield markers that hedges the speaker or writer and gives the degree of commitment to the proposition so that this person is protected in case his proposition proves wrong. Lemke's distinction between strong adjectives (e.g., critical and crucial) and week adjectives (appropriate and convenient) is not duplicable as the researcher tried to apply his model to the concordances of the extrapositions, found in both corpora, but unfortunately, it somehow did not work, so the researcher started interpreting the meaning of the different adjectives in the extrapositions intuitively as follows.

Looking closely at the Table 18, it can be easily discerned that NSs used more probability adjectives than NNSs. However, both used the same number of the adjective of importance. The first finding provides further evidence of the fact that NSs tend to show some kind of tentativeness in their writings.

5 Conclusion

Hedging, as a rhetorical or persuasion strategy gained a lot of popularity over the past 25 years and numerous scholars conducted studies on how the hedging devices can be used in academic writing [15, 50]. Two hedging taxonomies or models proposed by Hyland [8] and Salagar-Meyer [15] are applied to the two corpora. This application yielded the same result which is that NSs use more hedges than NNSs – 4,022 (19%) and 3,251 (12%), respectively. This finding is in line with Rezanejad, Lari and Mosalli's [42] study.

Although, generally, there is an overuse of modal auxiliaries by NNSs, some of these modals were mistakenly used. Similarly, Holmes [36] suggests that the overuse of the modal auxiliary *will* could be attributed to one of three hypotheses: either L1 transfer (interlingual); or developmental factor; or teaching-induced; or

speech-like writing (viz., students' writing is affected by the way they speak, i.e., register-interference).

It is interesting to notice that NSs use *may, could* and *would* (modals mainly expressing probability) more than NNSs. This could be attributed to the fact that NSs tends to use these modals when they want to show their attenuation about their propositions (epistemic stance). In other words, NSs prefer to use these probability modals when they give unproven truth in their proposition [50].

The second category of hedges (according to Hyland's model), employed by both native and non-native speakers, is the hedging adverbs. This time, the native speakers use more adverbs than non-native speakers; 1,096 and 723, respectively. Generally, within the global list of the hedging adverb, the difference between the two frequencies is statistically significant, particularly, the difference between the probability adverbs, such as '*perhaps*' which was used 81 times by NSs and eight times by NNSs. Similarly, adverbs like '*possibly', 'likely'*, '*roughly*' were greatly underused by the NNSs.

The finding of the underuse of hedging by non-native speakers was confirmed by Salagar-Meyer's [15] proposed taxonomy of hedging markers, which was applied to both corpora. The chi-square results suggest that there is a statistically significant difference between native and non-native speakers in their use of the hedging markers.

There seems to be unanimous agreement among researchers that NSs tend to 'downstate' while NNS tend to 'overstate' [51]. However, in this study and contrary to the expectations, NNS underused all scalar intensifiers (including both amplifiers and downtoners). This finding is in line with Granger [48]. The underpresentation of boosters in the non-native speakers' corpus is significant enough to be the cause of the underpresentation of amplifiers in general. However, the underuse of maximizers is ignored, as the difference is not significant.

The underuse of boosters is not confined to booster types but it includes the frequency of these boosters as well. When the lists of boosters compared, it is found that there are 22 types of boosters (with 38 frequencies in NSs' corpus) that are not deployed at all by the non-native speakers (e.g., *remarkably, desperately* and *agonizingly*). As mentioned before, this case of nonexistence of some boosters in the NNSs' corpus could be attributed to the "natural deficiency of non-native vocabulary" [32, p. 28]. Similarly, most of the boosters, underused by non-native speakers, were a combination of an intensifying adverb ending with the suffix -ly followed by an adjective (adv-adj-, e.g., increasingly difficult). This type of adverbial collocation requires high combinatory skill, which does not seem within the capabilities of the non-native speakers. To counteract this deficiency, NNSs resort to use all-round or stereotyped boosters that can be used in many contexts, such as *very*. This booster (i.e., very) is overused by NNSs, with 272 occurrences while NSs used it only 187 times. The non-native speakers' language deficiency is further corroborated by the lack of complex forms found in the native speakers' corpus such '*no more than'*. The core word of the previous phrase is the adverb '*more'*. This adverb in this context has been mistakenly classified by Wmatrix3 as a booster, but in reality and in this context, it is no more than a downtoner. If the

frequency of this adverb is taken away from the total of boosters in the NSs' corpus, this would reduce the number of amplifiers greatly. As mentioned earlier, NNSs do not only underuse the amplifiers, but the downtoners as well. The corroborating evidence for this underuse is found in the significant difference between the frequencies of downtoners in both NSs and NNSs' corpora (702 times and 414 times, respectively). Generally, downstating, as a way of hedging, is used to express vagueness and attenuation, which are two rhetorical strategies that distinguish a native speakers' writing [14, 49]. One of the important hedges that lies within the compromisers (subcategory of downtoners) is *sort of*. This hedge, which shows the degree of commitment of the writer towards the truth in a proposition, is significantly underused by the NNSs who were not trained or taught to exploit the indirect meaning of this hedge. Although NNSs underused this hedge, which shows the degree of commitment to the truth in their propositions, they overused the extraposition '*it… that'* which they used to indirectly comment on their propositions. The collected data suggests that the overuse could be attributed to a combination of factors. The substantive one is that they found this formulaic structure easy to start the sentence with. Furthermore, this structure is usually used to show some kind of objective modality and since there is difference in the quantity and quality between native and non-native speakers, this suggests that both groups use different ways to express modality [52]. Pedagogically, hedging is one of the areas that needs to be focused on by both language instructors and curriculum designers [35]. Data-driven learning (DDL), which is defined as the use of corpus concordances in classrooms, is one of the important applications of learner corpora.

References

1. Scarcella, R., Brunak, J.: On speaking politely in a second language. Int. J. Sociol. Lang. **1981** (1981)
2. Kasper, G.: Communication strategies: Modality reduction. Interlang. Stud. Bull. **4**, 83–266 (1979)
3. Robberecht, P., Petegham, M.: A functional model for the description of modality. In: The Fifth International Conference on Contrastive Projects (1982)
4. Naess, A.: Communication and Argument; Elements Of Applied Semantics. Allen and Unwin Limited, London (1966)
5. Skelton, J.: The care and maintenance of hedges. ELT J. **42**, 37–43 (1988)
6. Granger, S.: Modality in advanced Swedish learners' written interlanguage. In: Granger, S., Hung, J., Petch-Tyson, S. (eds.) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. John Benjamins Publishing Co, Netherlands (2002)
7. Scott, M., Tribble, C.: Textual Patterns. J. Benjamins, Philadelphia (2006)
8. Hyland, K.: Metadiscourse: Exploring Interaction in Writing (Continuum discourse series). Continuum International Publishing Group Ltd (2005)
9. Kopple, W.: Some Exploratory Discourse on Metadiscourse. Coll. Compos. Commun. **36**, 82 (1985)
10. Tse, P., Hyland, K.: So what is the problem this book addresses? Interactions in academic book reviews. Text & Talk—An Interdiscip. J. Lang. Discourse Commun. Stud. **26**, 767–790 (2006)

11. Dictionary, h.: Hedge meaning in the Cambridge English Dictionary. http://dictionary.cambridge.org/dictionary/english/hedge
12. Poos, D., Simpson, R.: Cross-disciplinary comparisons of hedging some findings from the Michigan Corpus of Academic Spoken English. In: Reppen, R., Fitzmaurice, S., Biber, D. (eds.) Using Corpora to Explore Linguistic Variation. John Benjamins Publishing Co, Amsterdam (2002)
13. Lakoff, G.: Hedges: A study in meaning criteria and the logic of fuzzy concepts. J. Philos. Logic **2** (1973)
14. McKenny, J.: A corpus-based investigation of the phraseology in various genres of written English with applications to the teaching of English for academic purposes (2006)
15. Salager-Meyer, F.: Hedges and textual communicative function in medical English written discourse. Engl. Specif. Purp. **13**, 149–170 (1994)
16. Murniato, M.: Types and functions of hedges used in 'J. K. Rowling's' interview with Oprah Winfrey show. http://kim.ung.ac.id/index.php/KIMFSB/article/download/3287/3263
17. Koutsantoni, D.: Developing Academic Literacies. Peter Lang, Oxford[etc.] (2007)
18. Quirk, R., Leech, G., Greenbaum, S., Crystal, D.: A Comprehensive Grammar of the English Language. Longman, London (1985)
19. Aijmer, K.: Modality in advanced Swedish learners' written interlanguage. In: Granger, S., Hung, J., Petch-Tyson, S. (eds.) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. John Benjamins Publishing Co, Netherlands (2002)
20. Hinkel, E.: The use of model verbs as a reflection of cultural values. TESOL Q. **29**, 325 (1995)
21. Papafragou, A.: Epistemic modality and truth conditions. Lingua **116**, 1688–1702 (2006)
22. Halliday, M.: An Introduction to Functional Grammar. Routledge, London (2014)
23. Haswell, R.: Gaining Ground in College Writing. Southern Methodist University Press, Dallas, Tex (1991)
24. Wray, A.: Formulaic Language and the Lexicon. Cambridge University Press, Cambridge (2002)
25. Lemke, J.: Resources for attitudinal meaning: evaluative orientations in text semantics. Funct. Lang. **5**, 33–56 (1998)
26. Corver, N., Van Riemsdijk, H.: Semi-lexical Categories: the Function of Content Words and the Content of Function Words. De Gruyter Mouton, Germany (2001)
27. Hinojosa, J., Martín-Loeches, M., Casado, P., Muñoz, F., Carretié, L., Fernández-Frías, C., Pozo, M.: Semantic processing of open- and closed-class words: an event-related potentials study. Cogn. Brain. Res. **11**, 397–407 (2001)
28. Corver, N., Van Riemsdijk, H.: Semi-lexical Categories. Mouton de Gruyter, Berlin (2001)
29. Halliday, M.: An Introduction to Functional Grammar. Hodder Arnold, London (2004)
30. Stubbs, M.: Text and Corpus Analysis: Computer-assisted Studies of Language and Culture. Blackwell, Oxford (1996)
31. Schmitt, N.: An Introduction to Applied Linguistics. Arnold, London (2002)
32. Lorenz, G.: Adjective Intensification Learners Versus Native Speakers: A Corpus Study of Argumentative Writing (Language & Computers). Rodopi, Amsterdam (1999)
33. Coventry University: British Academic Written English Corpus (BAWE). http://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/
34. Ortmeier-Hooper, C.: The ELL Writer. United States, United States (2013)
35. Hyland, K.: Hedging in academic writing and EAF textbooks. Engl. Specif. Purp. **13**, 239–256 (1994)
36. Holmes, J.: Doubt and certainty in ESL textbooks. Appl. Linguist. **9**, 21–44 (1988)
37. Kenny, A.: The Computation of Style: an Introduction to Statistics for Students of Literature and Humanities. Pergamon Press, Oxford (1985)
38. Scott, M.: WordSmith Tools, Version 3. Oxford (1999)

39. Ure, J.: Lexical density and register differentiation. In: Perren, J., Trim (ed.) Applications of Linguistics: Selected Papers of the 2nd International Congress of Applied Linguistics. Cambridge University Press, Cambridge (2017)
40. Meunier, F.: Computer tools for the analysis of learner corpora. Presented at the (1998)
41. Hunston, S., Francis, G.: Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English. Benjamins, Amsterdam [u.a.] (2000)
42. Rezanejad, A., Lari, Z., Mosalli, Z.: A cross-cultural analysis of the use of hedging devices in scientific research articles. J. Lang. Teach. Res. **6**, 1384 (2015)
43. Ringbom, H.: Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In: Granger, S., Leech, G. (eds.) Learner English on Computer. Longman, New York (2017)
44. Coates, J.: Semantics of the Modal Auxiliaries. Croom Helm, London (1983)
45. Henneren, L.: On Modality in English: A Study of the Semantics of the Modality. CWK Gleerup, Lund (1978)
46. Halliday, M.: Functional diversity in language, as seen from a consideration of modality and mood in English. Found. Lang. 6 (1970)
47. Archer, D., Wilson, A., Rayson, P.: Introduction to the USAS category system. http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf
48. Granger, S.: In: Cowie, A. (ed.) Phraseology: Theory, Analysis and Applications. Oxford University Press, Oxford (1998)
49. Channell.: Vague Language. Oxford University Press, Oxford (1994)
50. Hyland, K.: Nurturing hedges in the ESP curriculum. System **24**, 477–490 (1996)
51. Hyland, K., Milton, J.: Qualification and certainty in L1 and L2 students' writing. J. Second Lang. Writ. **6**, 183–205 (1997)
52. Johansson, M.: It-clefts and pseudo-clefts in Swedish advanced learner English. Moderna Sprak. **95**, 16–23 (2001)