Published in: Fuertes Olivera, P.A. (ed.) (2005). *Lengua y Sociedad: Investigaciones recientes en lingüística aplicada*. Lingüística y Filología no. 61. Valladolid: Universidad de Valladolid. 311-331.

Epistemic modality in MA dissertations

COSTAS GABRIELATOS AND TONY MCENERY
Universidad de Lancaster

1. INTRODUCTION

This paper reports on the compilation, and ongoing mark up and annotation, of a corpus of MA dissertations written by students at the Department of Linguistics and English Language, Lancaster University. The main focus of the paper is a preliminary investigation comparing the use of epistemic modality by native and advanced non-native speakers of English (henceforth NS and NNS respectively) in the corpus. The paper also outlines other possible uses of the corpus for research on academic English, made possible by the information encoded in the annotation and mark up.

2. POSITIONING THE STUDY

The compilation of learner corpora and analysis of learner language has proven helpful in research on learner errors, second language acquisition and the comparison between the language use of learners and native speakers (e.g. Granger, 1998b; Granger, Hung & Petch-Tyson, 2002). Corpus linguistics has also been interested in the analysis of academic discourse, often for purposes of academic writing instruction (e.g. Flowerdew, 2002). Hyland (1999) and Thompson & Tribble (2001) examined citation practices in published academic articles and PhD theses respectively. Hyland (2002) analysed the use of directives in academic writing using a corpus consisting of published articles and textbooks, as well as essays written by non-native speakers of English. Coxhead (2002) used a corpus of academic writing, consisting partly of sub-corpora from the LOB, Brown and Wellington corpora, to compile a word list for the teaching of academic English. Gledhill (2000) and Luzon Marco (2000) focused on frequent collocations in pharmaceutical and medical research articles respectively.

The availability of general and specialised corpora, as well as specialised sub-corpora within the large general ones containing native and non-native writing, makes it possible to carry out comparisons of specific linguistic features. A common feature of corpus research on both learner and academic language is the comparison to a norm. The language of learners, and non-native speakers in general, is compared to that of native-speakers, whose collective use as evidenced in corpora is taken as the norm, or at least as the target of learning[1] (e.g. Granger, 1997, 1998a; Virtanen, 1996). Similarly, the language practices of NS postgraduate and research students, who can be seen as trainee academics, is compared to that of established writers as evidenced in the discourse of published papers (e.g. Thompson, 2002). An approach sharing features of the previous two is the comparison of academic articles, irrespective of whether they were written by native or non-native speakers, on the basis of specific features (e.g. Lucas et al., 2003). In all the above cases the comparison is two-way, between 'student'/'novice' and 'expert' in terms of language use in general, or academic writing in particular (see figure 1).

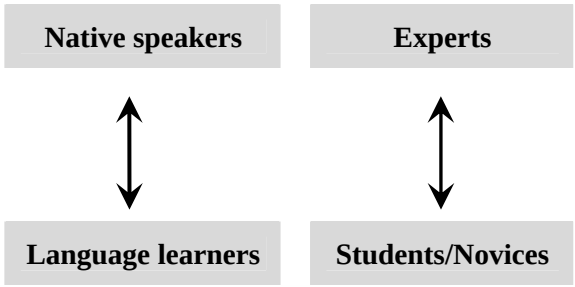

Figure 1. Two-way comparison in corpus studies

Of course this polar relationship is a simplification of the degrees that can be distinguished. Language competence can be graded from beginner to native speaker, as is the case in English language teaching. Expertise in academic writing may be graded, as is the usual practice, according to educational levels: secondary school, undergraduate, graduate and research student writing.

The MA corpus used in this paper, containing the same text type written by NS and NNS, if supplemented by a corpus of established academic writing in linguistics and language teaching, lends itself to a combination of the two approaches, adding a further dimension (see figure 2).

[1] This is particularly evidenced in the error-tagging of learner corpora (e.g. Granger, 1999).

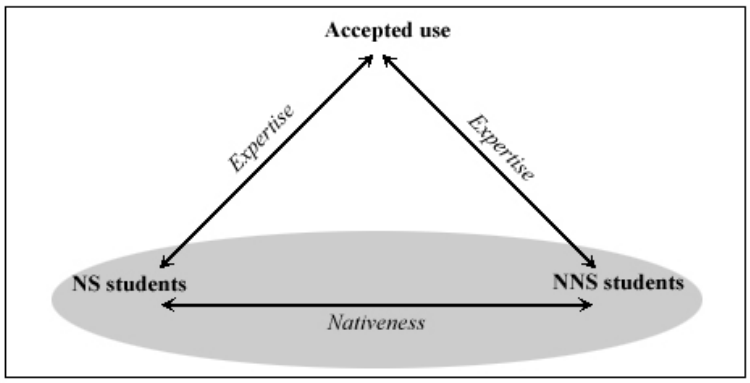

Figure 2. Three-way comparison in corpus studies and positioning of the present study

Arguably, the NNS students in the corpus used in this paper are at an advanced level as far as their general language competence is concerned, as they have been accepted to read for a higher degree using English as both the medium of instruction and assessment. However, they are also continuing learners of English, whose language use can be compared with that of NS. At the same time, both NNS and NS students are learners of academic writing in general, and research papers in particular, and their writing can be compared with that of experts.[2] As figure 2 shows, the students' academic writing can be compared on the basis of both 'nativeness' and 'expertise'. This study concerns itself with a comparison in terms of the former element (represented by the shaded area in figure 2).

3. WHY EXAMINE EPISTEMIC MODALITY IN MA DISSERTATIONS?

Epistemic modality is concerned with the expression of the users' degree of certainty, or commitment to the truth, of their statements, or the assessment of the likelihood of something being, or having been, the case (e.g. Biber et al., 1999; Coates, 1983; Huddleston & Pullum, 2002; Palmer, 1986, 1990; Quirk et al., 1985). A small number of auxiliary verbs, also termed modal auxiliaries (*can, could, may, might, must, shall, should, will, would*), are regarded as the prototypical morphological realisation of epistemic modality (e.g. Huddleston,

[2] Interestingly, if acceptable academic writing is defined according to selection for publication, then the published writing of NNS should also be included in a corpus of expert academic writing. Although this issue may need to be resolved in further studies using the MA corpus, it falls outside the scope of the present paper.

1984: 164). However, epistemic modality is also expressed by a number of lexical verbs (e.g. *believe, infer, know*), adjectives (e.g. *definite, probable, unlikely*), adverbs (e.g. *arguably, certainly, possibly*), and multi-word units and colligations involving lexis expressing degrees of certainty (e.g. *call into question, chances are, it seems plausible*).

Academic writing lends itself to studies of epistemic modality, as one of the characteristics of the genre is the frequent use of hedging (e.g. Thompson, 2002). Consequently, the ability to qualify statements appropriately (e.g. conveying the degree of certainty that the research evidence calls for) is central to good academic writing and is, therefore, a skill that MA students, NS and NNS alike, need to master.

The examination of the degree of certainty expressed in the dissertations, as well as the selection of linguistic devices in terms of types and frequency can provide helpful indications as to the areas MA students need help with, which can then form the basis for support seminars and workshops before students begin work on their dissertations. The comparison between native and non-native writers, as well as between writers of different native languages, can yield interesting patterns as to the problematic areas that are specific to each group (e.g. Hyland & Milton, 1997; McEnery & Kiefle, 2002).

3.1 The corpus

The corpus consists of 273 MA dissertations produced by postgraduate students studying at the Department of Linguistics and English Language, Lancaster University. At the time of this study, the corpus consisted of the main body of the dissertations; that is, titles, acknowledgements, abstracts, reference lists and appendices were excluded from the corpus.[3]

In terms of first language, 85 students (31.1%) were NS, and 188 (68.9%) were NNS. The corpus comprises 3,790,661 words, of which 1,146,864 (30.26%) were written by NS and 2,643,797 (69.74%) were written by NNS. The vast majority (79, 92.9%) of NS were speakers of British English. Three languages are dominant in the NNS sub-corpus, representing more than half (55%) of the dissertations in the sub-corpus: Chinese (48)[4], Greek (34)[5], and Japanese (22). Table 1 shows the breakdown of the NNS students in terms of country of origin.

[3] See section 6 for an outline of the ongoing annotation of quotes, examples of data, statistics etc.
[4] The figure includes students from China, Hong Kong and Taiwan.
[5] The figure includes students from Greece and Cyprus.

Country	No. of students
Greece	*32*
China, Japan	*22*
Hong Kong	*15*
Malaysia	*13*
Taiwan	*11*
Korea	*8*
Italy	*6*
Russia, Spain	*5*
Botswana, Brazil, Germany	*3*
Belgium, Burma, Colombia, Cyprus, Eritrea, India, Malawi, Nepal, Nicaragua, Thailand	*2*
Angola, Austria, Chile, Denmark, Estonia, France, Holland, Indonesia, Jamaica, Kenya, Luxemburg, Mexico, Morocco, Mozambique, Oman, Singapore, Sudan, Tanzania, Tunisia, Turkey	*1*

Table 1: Country of origin of NNS students

The two sub-corpora (NS and NNS) are comparable in many respects. The texts belong to the same genre and the topics revolve around issues in linguistics and English language learning and teaching.[6] In other words, all students, NS and NNS alike, were striving to conform to the linguistic practices of the same discourse community, and were all adhering to the same broad organisational framework. All writers were students in the same department, and the texts span a short period of time (96% of the dissertations were submitted between 1996 and 2002 and the rest between 1992 and 1993).

As expected, the MA dissertations in the corpus contain direct quotes from the literature, excerpts from data (e.g. questionnaire answers, transcribed interviews, excerpts from pedagogical materials), as well as non-prose text (e.g. tables, charts, graphs). Arguably, including those non-author and non-prose (henceforth NANP) text portions in the analysis would skew the overall results, and would probably obscure differences between native and non-native students, as well as among speakers of different first languages. To avoid this, the non-prose portions have been removed from the corpus, and the non-author portions are in the process of being annotated (see section 6). However, it should also be noted that informal observations during the manual NANP annotation indicate

[6] The students had followed the MA Language Studies or MA Language Teaching programmes.

that non-author text constitutes a relative low percentage of the total number of words in the dissertations, and, therefore, the version of the corpus used in the preliminary study reported here can yield dependable results regarding the use of many epistemic expressions by the NS and NNS students in the corpus. However, since informal observations, not unlike intuitions, are not always dependable, it would be interesting to check them against a comparative word-count of the fully NANP annotated corpus.

Given that the dissertations also contain summaries and paraphrases of views in the literature, it can be argued that these passages contain, or are strongly influenced by, the use of modality in the original texts. However, this is not expected to have influenced the results of this study unduly for two interrelated reasons. Firstly, the use of modality in those passages is expected to have been affected by the students' interpretation of the source text. Secondly, the students' overall writing style, and hence what this study treats as 'their' use of epistemic modality, has arguably been shaped in the first place by the academic texts they have read and are presumably emulating.

The preliminary examination reported in this paper is based on the first version of the corpus, which contains non-author and non-prose text. Because, as was mentioned above, findings may be skewed because of the non-author portions of text (quotes and data), this study does not carry out a separate examination of the use of epistemic modality by the two groups, nor a comparison with a representative corpus. Rather, it is restricted to a comparison between the NS and NNS sub-corpora. The comparison of the use of epistemic modality between NS and NNS students is expected to yield reliable results, because, arguably, the presence of non-author text portions in both sub-corpora cancels out the effect of one another to a significant degree. However, any figures should be treated more as comparative indicators than as reliable in their own right.

3.2 Annotation and analysis

The fact that epistemic modality is realised by a large number of words and multi-word units requires that the corpus be annotated semantically as well as morphologically. Morphological and semantic annotation were carried out automatically by the Wmatrix tool (Rayson, 2001, 2003). The part-of-speech annotation used the CLAWS tagset (Garside, 1987), and the semantic annotation used the USAS category system (Archer et al., 2002). The USAS system categorizes meaning according to broad semantic fields, e.g. "Terms relating to reasoning/thinking, and level of belief/scepticism" (ibid.: 31). In total, USAS annotates using 232 category labels.

As the USAS categories were developed with much broader applications in mind, epistemic expressions are found in five different categories (A7, X2.1, X2.2, X2.6, and T1.1.3). Category A7 includes the majority of epistemic expressions, with epistemic *will* being the most notable omission. Epistemic *will* is subsumed with futurity *will* in category T1.1.3. Excepting A7, these categories also include items which do not express epistemic modality. Finally, epistemic *must* and *should* are not covered by any category. Consequently, the automatic semantic annotation has to be supplemented with the manual annotation of the relevant items before the full analysis can be carried out. Table 2 provides descriptions of the categories related to epistemic modality, examples of items that fall within the scope of the study, and examples of items that need to be manually excluded from the analysis, as they do not express epistemic modality.

Code	**Description**	**Examples of epistemic expressions**	**Items expressing other concepts**
A7 Definite (+modals)	Abstract terms of modality (possibility, necessity, certainty, etc.)	Modal auxiliaries • *can, could, may, might, would.* Modal lexis • *achievable, certain, positive, possible, potential, probable, tentative, by all means, grey area, have a chance, no matter what, no two ways about it, out of the question*	
X.2.1 Thought, belief	Terms relating to reasoning/thinking, and level of belief/scepticism	*assume, believe, presumably*	*conceptualise, formulate, images*
X2.2 Knowledge	Terms relating to (level of) knowledge/ perception/retrospection	*anybody's guess, can't tell*	*acquainted, cognisant, forget, hindsight*
X2.6 Expect	Terms depicting (level of) expectation	*anticipate, forecast, foresee*	*ironically, on impulse, out of the blue*
T1.1.3 Time: General: Future	General terms relating to a future (period/point in) time	*gonna, shall, will*	*defer, future, postpone, tomorrow*

Table 2. Semantic categories (wholly or partly) relevant to epistemic modality

The preliminary analysis concerns itself with epistemic expressions included in category A7 (see table 2 above) for three reasons. The category is by far the most populous for epistemic expressions, it contains only epistemic expressions, and includes five central modals. For the reasons outlined above, a more detailed treatment is reserved for the comparison of the ten most frequent A7 epistemic expressions in the NS and NNS sub-corpora.

4. FINDINGS AND COMMENTS

The two groups show virtually the same high concentration of use of A7 types. In both sub-corpora, the ten most frequent A7 types account for more than 80% of the A7 tokens, despite representing a mere 6% of the total A7 types in either sub-corpus (see table 3).

NS	*6.2%*	of A7 types account for	*82.9%*	of A7 tokens in the sub-corpus
NNS	*6.0%*		*83.1%*	

Table 3. Representation of the 10 most frequent A7 types in the sub-corpora

The concentration is made clearer if we consider the five most frequent A7 types, which, perhaps unsurprisingly, are the five central modals included in the A7 category (*can, could, may, might, would*). These five modal auxiliaries account for some 70% of the A7 tokens in the sub-corpora, although they only represent 3% of the total A7 types (see table 4).

NS	*3.1%*	of A7 types account for	*68.7%*	of A7 tokens in the sub-corpus
NNS	*3.0%*		*70.5%*	

Table 4. Representation of the 5 most frequent A7 types in the sub-corpora

The tight clustering of A7 use around a very small number of items in both sub-corpora is made more evident when types are lemmatised.[7] The ten most frequent A7 lemmas account for some 21% of the A7 types and about 90% of the A7 tokens (see table 5). If we compare the figures in tables 3 and 5, we realise that an almost four-fold increase in the number of types only yields an increase of under 7% in the proportion of tokens .

NS	*21.1%*	of A7 types (representing the top 10 lemmas) account for	*90.4%*	of A7 tokens in the sub-corpus
NNS	*21.6%*		*89.5%*	

Table 5. Representation of the 10 most frequent A7 lemmas in the sub-corpora

Even more striking than the close similarity in the clustering of epistemic use around a very small number of expressions is the almost identical make-up of the 10 most frequent A7 types in the two subcorpora. Not only do the two groups share nine out of the ten most frequent A7 types, but also eight out of ten, including the top six, are in the same order of frequency. Another significant similarity is that in both cases the five most frequent A7 expressions are modal auxiliaries, and the next five are adjectives and adverbs (see table 6).

Order of frequency	**NS**	**NNS**
1	*can*	*can*
2	*would*	*would*
3	*may*	*may*
4	*could*	*could*
5	*might*	*might*
6	*possible*	*possible*
7	*perhaps*	*positive*
8	*likely*	*likely*
9	*positive*	*probably*
10	*potential*	*potential*

Table 6. The top ten epistemic expressions in the sub-corpora

[7] There is controversy as to whether the central modals should be approached as the present/past forms of the same root (i.e. *can/could, may/might, shall/should, will/would*), or as separate entities (e.g. Leech, 2004: 141-143), and, consequently, as to whether lemmatisation of modal auxiliaries is acceptable. As it is not within the scope of this paper to resolve this controversy, lemmatisation was treated as an opportunity to examine frequent roots that produce lexis expressing epistemic modality (e.g. possible/possibly/possibility).

Also, in both the NS and NNS groups, the five modal auxiliaries (*can, could, may, might, would*) account for the majority of A7 tokens in the respective sub-corpora (see table 7).

	NS	**NNS**
can, could, may, might, would	*68.7%*	*70.5%*
likely/probably, perhaps, possible, positive, potential	*14.2%*	*12.6%*

Table 7. Proportion of tokens corresponding to the top ten A7 types in relation to the total number of A7 tokens in the sub-corpora.

Lemmatisation clusters the ten most frequent A7 expressions more tightly, reveals a few more frequent items, and strengthens the similarity between the ten most frequent A7 expressions[8] used by the NS and NNS students in the corpus. Again, nine out of the ten most frequent A7 lemmas[9] are common in the two sub-corpora, and six share the same ranking (see table 8).

Rank	**NS**	**NNS**
1	CAN	CAN
2	MAY	MAY
3	would / ‘d	would / ‘d
4	POSSIBLE	POSSIBLE
5	perhaps	POSITIVE
6	LIKELY	LIKELY
7	POSITIVE	POTENTIAL
8	POTENTIAL	PROBABLE
9	CERTAIN	CERTAIN
10	SURE	perhaps

Table 8. The ten most frequent A7 lemmas in the sub-corpora

[8] Despite the absence of epistemic *must*, *should* and *will* from the results the table provides a good overall picture of the most frequent epistemic expressions.

[9] Lemmas are indicated by capital letters. *Perhaps* and *would* are treated as types, the former because it is not productive, the latter because the analysis does not include *will*.

However, the striking similarities in the make-up and ranking of the ten most frequent A7 types and lemmas hide some very interesting differences. As shown in table 9, the NS students use epistemic expressions much more frequently than the NNS (approximately 50% more often).

	NS	**NNS**
Number of words in the sub-corpus	*1,146,864*	*2,643,797*
Frequency of A7 tokens	*16,143*	*24,412*
Frequency of A7 tokens per 1,000 words	*14.07*	*9.23*

Table 9. Frequency of A7 tokens in the sub-corpora

In terms of the relative variety of epistemic expressions, the NS sub-corpus may contain slightly fewer A7 types than the NNS one (161 and 167 respectively), but this can be attributed to the NS sub-corpus being about half the size of the NNS one. Given the relative size of the two sub-corpora, the fact that they exhibit almost the same number of A7 types is a good indication that the NS students use a wider range of A7 epistemic expressions. This is confirmed by the comparison of the type-token ratios of A7 items in the NS and NNS sub-corpora, 1% and 0.68% respectively (see table 10). Interestingly, the same relation between the frequency of A7 tokens also holds between the type-token ratios, that is, NS students use A7 tokens about 50% more frequently, and, proportionately, about 50% more A7 types than the NNS students.[10]

	NS	**NNS**
Frequency of A7 tokens	*16,143*	*24,412*
A7 types	*161*	*167*
A7 type-token ratio	*1%*	*0.68%*

Table 10. Frequency of A7 tokens and A7 type-token ratio in the sub-corpora

[10] This figure is expected to be higher when the non-author portions have been removed from the calculations, as informal observations during the manual annotation of non-author portions have so far indicated that NNS students tend to make use of quoting more often than the NS students. That is, part of the frequency and variety of A7 items shown in this preliminary study is expected to be attributed to the quotes from the literature.

More differences are revealed if we examine the frequencies of the ten most common A7 types as a proportion of the number of words in each sub-corpus, that is, their frequency per 1,000 words in the respective sub-corpora. A distinct difference is highlighted when we compare the relative frequencies of the modal auxiliaries and the adjectives/adverbs within the ten most frequent types. The NS students use A7 modal auxiliaries just under 50% more frequently than the NNS students, whereas they use A7 adjectives/adverbs almost 75% more frequently (see table 11 below).

	Frequency per 1,000 words	
	NS	**NNS**
Modal auxiliaries	*9.66*	*6.51*
Adjectives and adverbs	*2.16*	*1.25*

Table 11. Comparison of the frequency of modal auxiliaries and adjectives/adverbs in the sub-corpora

This marked difference in the use of epistemic adjectives and adverbs by NS and NNS students was obscured when we considered the top ten A7 items as one group (see table 9), because the A7 modal auxiliaries in both sub-corpora are almost five times more frequent than the A7 adjectives and adverbs together. The reasons for the relatively low frequency of epistemic adjectives and adverbs in the NNS sub-corpus can be traced to pedagogical materials, in which discussion and examples of modality tend to revolve around the central modals.

Let us now turn out attention to the comparison of the frequencies per 1,000 words of the ten most frequent types in each sub-corpus.[11] In the majority of cases (eight out of eleven) the relative frequencies are consistent with the general picture, that is, NS students use the A7 types more frequently than NNS (see tables 10 and 11 above). There is only one A7 type which the two groups use with the same frequency (*positive*), and two (*can, probably*) which, although slightly more frequent in the NS sub-corpus, should perhaps be treated as showing comparable frequencies (see figure 3 and table 12).

[11] The total number of types to be compared is 11. As the two groups share nine out of the ten most frequent types (see table 6) we need to add the two that are not common to the two sub-corpora (*perhaps* and *probably*).

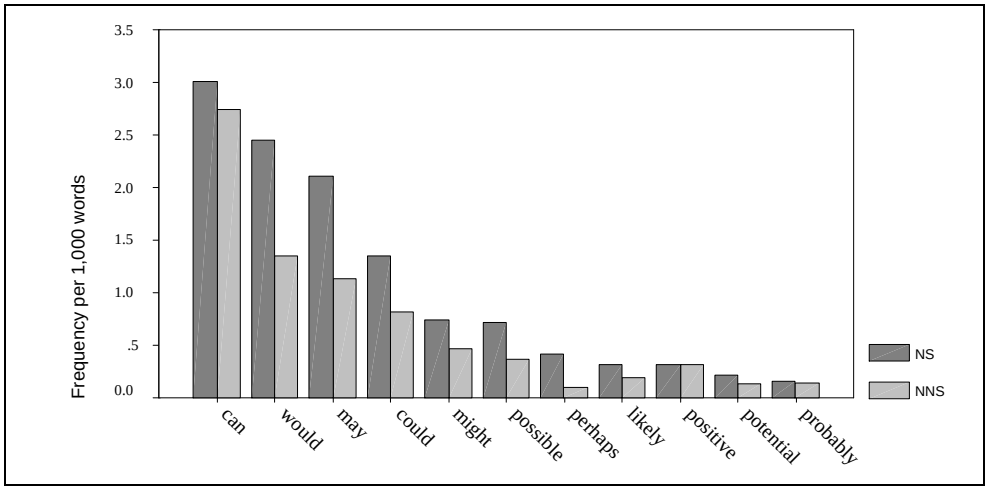

Figure 3. Comparison of the frequencies per 1,000 of the most frequent A7 types in the two sub-corpora

When comparing the relative frequencies of the most frequent A7 types it is advisable to take into account the log-likelihood statistic to ensure that any similarities and differences are statistically significant,[12] as the two sub-corpora are of different sizes (see table 12).

	Frequency per 1,000 words		**LL**
	NS	NNS	
would	*2.45*	*1.35*	536.48
may	*2.11*	*1.13*	499.61
perhaps	*0.42*	*0.10*	359.68
could	*1.35*	*0.82*	211.09
possible	*0.72*	*0.37*	187.87
might	*0.74*	*0.47*	85.58
likely	*0.32*	*0.19*	60.18
potential	*0.22*	*0.13*	34.17
can	*3.01*	*2.74*	21.76
probably	*0.16*	*0.14*	0.98
positive	*0.32*	*0.32*	0.05

Table 12. Comparison of the top ten A7 types, ordered by descending log-likelihood values

[12] The cut-off value for statistical significance at the 0.01% level used in this study is 15.13 (see Rayson et al., 2004).

Apart from the cases of *positive* and *probably*, the similarities and differences in the use of A7 items by the two groups are significant at least at the 0.01% level. The most marked differences in the relative frequency of use, that is, the cases where frequency of use by the NS students is much higher than the 50% average, are mainly (in decreasing order) *perhaps, possible, may, would,* and to a lesser degree *likely, potential, could* (see table 13).

	Difference NS/NNS (%)	**LL**
perhaps	+320%	359.68
possible	+94%	187.87
may	+86%	499.61
would	+81%	536.48
potential	+69%	34.17
likely	+68%	60.18
could	+64%	211.09
might	+57%	85.58
probably	+14%	0.98
can	+10%	21.76
positive	0%	0.05

Table 13. Comparison of the top ten A7 types, ordered by descending % difference

A first observation is that the differences do not seem to show any particular bias towards modal auxiliaries or adjectives/adverbs. It must also be noted that of the three types that are used as frequently per 1,000 words by both groups (*can, positive, probably*) only *can* registers statistical significance. This observation suggests that the differences in the use of epistemic modality between the NS and NNS groups are stronger than the similarities.

The most pronounced difference by far is in the use of *perhaps*, with a frequency more than four times higher in the NS sub-corpus. One possible explanation for this is that English language teaching materials treat *perhaps* as being rather informal. For example, the Longman Dictionary of Contemporary English (2003: 1221) offers the following usage note: "*May* or *might* usually sounds more natural than *perhaps ... will*".[13]

[13] Our italics - original is in boldface.

5. CONCLUSIONS

The preliminary comparison of the use of epistemic modality by NS and NNS MA students in linguistics and language teaching, as represented in the use of A7 lexis in the corpus, shows interesting similarities and differences in usage. In both groups, epistemic use clusters tightly around a small number of types, accounting for the vast majority (some 83%) of A7 tokens. What is more, the two groups share nine out of the ten most frequent types, in almost identical order of frequency. However, the comparison of the frequency of A7 tokens per 1,000 words and the type-token ratios of the top ten types in the two groups reveal some important differences. Overall, NS students use A7 expressions much more frequently than NNS (+50%). The difference is even more distinct in the case of epistemic adjectives/adverbs, which NS students use about 75% more frequently than NNS.

At this point, it is difficult to determine the reasons behind the differences, which, arguably, are made all the more intriguing by the similarities. The differences in the frequency of epistemic expressions observed in the MA corpus are compatible with the findings of Hyland & Milton (1997), who concluded that the Chinese secondary school students in their study demonstrated a higher degree of assertiveness, or commitment to their statements, than the NS students. However, McEnery & Kifle (2002) observed the opposite trend in their corpus. That is, Eritrean secondary school students used weak epistemic modals more frequently than NS students, a trend which the researchers tentatively attributed to the influence of the instructional materials, as the use of epistemic modality by the learners in the study reflected the information given in the textbooks used in Eritrea.

Overall, the reasons for any similarities or differences between the NS and NNS students in the corpus are expected to be traced, at least partly, to the first language and/or the cultural and educational contexts of the NNS students. That is, the use of epistemic modality by NNS students is expected to be influenced by the status and practices of epistemic modality in their culture and first language in general, and in their educational and academic contexts in particular. Also, it is not unlikely that, to some extent, the explanation for the interesting similarities and differences rests on the tension between, on the one hand, the information that NNS students received in their English instruction, and, on the other, the discourse of the academic materials they read during their English-medium studies.

However, these tentative suggestions should be read against certain reservations. In this paper, non-native speakers are treated as a homogeneous group, an approach which may obscure differences between the epistemic use of students with different first languages and/or cultures. The differences may be due to two

factors, operating either individually or in combination: interference from the L1 and the influence of instruction. Another parameter which has not been taken into consideration in this study is the topic of each dissertation. A study of the use of modality in PhD theses (Thompson, 2002) revealed differences between disciplines. It is possible, then, that a comparison of the use of epistemic modality in terms of both first language and dissertation topic will reveal a more complex picture.

6. FURTHER STEPS

The detailed study will distinguish between the use of epistemic modality by speakers of different first languages, and for that purpose information about the students' first language has been added to the header of each dissertation in the corpus.[14] The header also contains a number of useful meta-data about each dissertation, which makes the corpus a more versatile research tool.

For example, the header contains information about the grade each dissertation was awarded as well as the final grade for each student. This enables correlations between the students' use of specific linguistic features and their academic achievement. It also encodes information about the area and topic of the dissertation. This information enables the addition of a further dimension of comparison, one between dissertations focusing on issues in linguistics or language teaching. Table 14 gives an example of a header and explains the nature of information included.

HEADER META-DATA	
Fields and examples	**Explanation**
<id="002"> </id>	Reference number of dissertation after anonymisation
<course="MALT"> </course>	The particular MA course the student completed. In this case it is the MA in Language Teaching.
<L Eng="FL" L1="Dutch"> </L>	The status of English for the student, either foreign language (FL) of mother tongue (MT).
<grade D ="4" M="2.5"> </grade>	The grades awarded to the dissertation and the Master's respectively.

[14] The addition of the header information was carried out after the completion of the preliminary analysis. The authors would like to gratefully acknowledge the contribution of Dr. Alan Waters (Lancaster University) to the mark-up process.

HEADER META-DATA	
Fields and examples	**Explanation**
<score result="B" exam="cambP" alte ="5"> </score>	This field is only relevant to non-native speakers. It provides information about the language examination score on the basis of which the student was admitted. In this case the student was admitted on the strength of a 'B' grade in the Certificate of Proficiency in English (UCLES - now Cambridge ESOL). The scores of the different examinations are normalised with reference to the levels of the Association of Language Testers in Europe.
<degree level="MA" area="Critical Discourse Analysis" topic="Educational Policy, Bilingualism"> </degree>	This field includes information about the level of the degree (as the corpus may be expanded to include PhD theses), the general area of the dissertation, and the topic(s) examined in the dissertation.

Table 14. Outline of header information

As was mentioned in 3.2, the corpus is being annotated for the text portions that are not written by the author (e.g. quotations), and text in non-prose form (e.g. tables). The information encoded in the non-author, non-prose (NANP) annotation refers to their content, source and format (for an outline see tables 15, 16 and 17)

NANP: Content		
Description	**Tag**	**Details**
Quote	Q	Prose excerpt from the literature.
Example	E	Not attested language example.
Statistics	S	Numerical presentation of findings or analysis.
Question	QQ	From questionnaire or interview.
Rubric	R	Instructions or questions from tests, exercises, tasks etc.
Term	TR	When cited or containing target element (in our case, terms or lexis related to modality).
Outline	O	Elements or breakdown of a test, framework etc. Sequence of research or lesson stages. Summary in table form. It may be presented in verbal and/or visual form.
Data	D	Text that is the object of analysis: corpus, press/media, published text, pedagogical materials (example, text, excerpt, rule, exercise etc.), elicited response (questionnaire, interview, test, assignment), field data (recording of spontaneous speech).

Table 15. Annotation scheme for the content of non-author and non-prose text.

NANP: Source		
Description	**Tag**	**Details**
Author	A	Non-attested language example, term.
Literature	L	Book, article, webpage, reference book etc.
Corpus	C	
Questionnaire	QR	Used by the author or cited from the literature
Interview	I	Carried out by the author or cited from literature.
Recording	R	Discussion, conversation etc. not structured by researcher.
Pedagogical materials	PM	Excerpts from grammars, dictionaries, coursebooks, or textbooks for language learners, used as data or examples.
Test	TS	Anything written/spoken by learners in exam conditions.
Home Assignment	HA	Texts not written under exam conditions.
Text	TX	Text from the public domain used as data.

Table 16. Annotation scheme for the source of non-author and non-prose text.

NANP: Format		
Description	**Tag**	**Details**
Prose	P	Includes dialogue.
Concordance	CC	Corpus data in concordance format.
List	LL	
Table	TB	
Graph	G	Mainly visual representation, with labels (e.g. pie chart, bar chart, line).
Diagram	D	Combination of visual representation and verbal information.

Table 17. Annotation scheme for the format of non-author and non-prose text.

The NANP annotation further expands the possible uses of the corpus. One of the next steps will be to compare the use of epistemic modality in the quotations and in the students' writing. Of course, the comparison can focus on any other linguistic feature. The fully annotated corpus will also lend itself to research on students' academic practices (e.g. Thompson, 2000). For example, the focus of investigation may be the proportion of direct citations to the total number of words for NS and NNS students, or the way quotations are incorporated in the text.

7. REFERENCES

AARTS J., DE MÖNNINK, I. & WEKKER, H. (eds.) 1997. *Studies in English Language and Teaching.* Amsterdam & Atlanta: Rodopi.

ARCHER, D., WILSON, A. & RAYSON, P. 2002, October. "Introduction to the USAS Category System." Benedict project report. Available online: http://www.comp.lancs.ac.uk/ucrel/usas/usas%20guide.pdf

ARCHER, D., RAYSON, P., WILSON, A. & MCENERY, T. (eds.) 2003. *Proceedings of Corpus Linguistics 2003.* Technical Papers Vol. 16 - Special Issue, University Centre for Computer Corpus Research on Language, Lancaster University.

BIBER, D., JOHANSSON, S., LEECH, J., CONRAD, S. & FINEGAN E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

BURNARD, L. & MCENERY, T. (eds.) 2000. *Rethinking Language Pedagogy from a Corpus Perspective. Papers from the third international conference on Teaching and Language Corpora*. Frankfurt am Main: Peter Lang.

COXHEAD, A. 2002. "The Academic word list: a corpus-based word list for academic purposes." In Kettemann, B. & Marko, G. (eds.), 73-89.

FLOWERDEW, J. (ed.) 2002. *Academic Discourse.* Harlow: Longman.

GARSIDE, R. 1987. "The CLAWS word-tagging system." In Garside, R. et al. (eds.), 30-41.

GARSIDE, R., LEECH, G. & SAMPSON, G. (eds.) 1987. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

GLEDHILL, C. 2000. "The discourse function of collocation in research article introductions." *English for Specific Purposes 19*, 115-135.

GRANGER, S. 1997. "On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared." In Aarts J. et al. (eds.), 185-198.

— 1998a. "The Computer Learner Corpus: a versatile new source of data for SLA research." In Granger, S. (ed.), 3-18.

— (ed.) 1998b. *Learner English On Computer*. London: Longman.

— 1999. "Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus." In Hasselgard, H. & Oksefjell, S. (eds.), 191-202.

GRANGER, S., HUNG, J. & PETCH-TYSON, S. (eds.) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

HASSELGARD, H. & OKSEFJELL, S. (eds.) 1999. *Out of Corpora. Studies in Honour of Stig Johansson.* Amsterdam: Rodopi.

HUDDLESTON, R. 1984. *Introduction to the Grammar of English*. Cambridge: Cambridge University Press.

HUDDLESTON, R. & PULLUM, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

HYLAND, K. 1999. "Academic attribution: citation and the construction of disciplinary knowledge." *Applied Linguistics 20*(3), 341-367.

— 2002. "Directives: Argument and engagement in academic writing." *Applied Linguistics* 23(2), 215-239.

HYLAND, K. & MILTON, J. 1997. "Qualification and certainty in L1 and L2 students' writing." *Journal of Second Language Writing* 6(2), 185-205.

KETTEMANN, B. & MARKO, G. (eds.) 2002. *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi.

LEECH, G. 2004. "A new Gray's anatomy of English grammar." *English Language and Linguistics* 8(1), 121–147.

LJUNG, M. (ed.) 1996. *Corpus-based Studies in English.* Papers from the seventeenth International conference on English Language Research on computerized Corpora (ICAME 17). Amsterdam: Rodopi.

Longman Dictionary of Contemporary English (4th ed.), 2003.

LORENZ G. 1999. *Adjective Intensification - Learners versus Native Speakers. A Corpus Study of Argumentative Writing*. Language and Computers: Studies in Practical Linguistics 27. Amsterdam & Atlanta: Rodopi.

LUCAS, N., CRÉMILLEUX, B. & TURMEL, L. 2003. "Signalling well-written academic articles in an English corpus by text mining techniques." In Archer, D. et al. (eds.), 465-474.

LUZON MARCO, M.J. 2000. "Collocational frameworks in medical research papers: A genre-based study." *English for Specific Purposes 19*, 63-68.

LYONS, J. 1977. *Semantics.* Cambridge: Cambridge University Press.

MCENERY, T. & KIFLE, N.A. 2002. "Epistemic modality in argumentative essays of second-language writers." In Flowerdew, J. (ed.), 182-195.

PALMER, F. R. 1986. *Mood and Modality*. Cambridge: Cambridge University Press.

— 1990. *Modality and the English Modals*. London: Longman.

PURNELLE, G., FAIRON, C. & DISTER, A. (eds.) 2004. *Le Poids des Mots.* Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004), Vol.2, Louvain-la-Neuve, Belgium, March 10-12, 2004, Presses Universitaires de Louvain.

QUIRK, R., GREENBAUM, S., LEECH, G. & SVARTVIK, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

RAYSON, P. 2001. "Wmatrix: a web-based corpus processing environment." Software demonstration presented at ICAME 2001 Conference, Université Catholique de Louvain, Belgium, May 16-20. Available online: http://www.comp.lancs.ac.uk/computing/users/paul/publications/icame01.pdf.

RAYSON, P. 2003. "Matrix: a statistical method and software tool for linguistic analysis through corpus comparison." Unpublished PhD thesis, Lancaster University. Available online:http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf.

RAYSON, P., BERRIDGE, D. & FRANCIS, B. 2004. "Extending the Cochran rule for the comparison of word frequencies between corpora." In Purnelle, G. et al. (eds.), 926-936.
THOMPSON, P. 2000. "Citation practices in PhD theses." In Burnard, L. & McEnery, T. (eds.), 91-101.
— 2002. "Modal verbs in academic writing." In Kettemann, B. & Marko, G. (eds.), 305-325.
THOMPSON, P. & TRIBBLE, C. 2001. "Looking at citations: using corpora in English for Academic Purposes." *Language Learning & Technology* 5(3), 91-105.
VIRTANEN, T. 1996. "The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English." In Ljung, M. (ed.), 299-309.