Šarolta Godnič Vičič

Potentials and Challenges of ESP Learner Corpora: The Case of Modal Auxiliaries in Slovene ESP Learners' Written Interlanguage

Abstract

A corpus-based approach to interlanguage analysis has been used for over a decade now and its impacts on foreign language teaching materials and teaching practices has been quite substantial. The approach, however, has rarely been used for studying ESP learners' interlanguage. This paper therefore aims to address the potentials that a small ESP learner corpus can present to ESP teachers, and determine whether the resources available can at present support such an analysis. Based on a small corpus of ESP student essays, central modal auxiliaries were studied with a focus on overuse errors. The analysis revealed that providing a reliable explanation for students' overuse errors is rather difficult given the existing corpus resources and that these are still too few in the fields of ESP. Nevertheless, the results can provide sufficient grounds for ESP teachers to adjust their teaching materials to their learners' needs.

Keywords: interlanguage, overuse errors, modal auxiliaries, ESP learner corpora.

Introduction

Research into second language acquisition (SLA) has shown that interlanguage - i.e., language produced by learners of a foreign language while trying to master it - is affected by a number of linguistic, psycholinguistic and situational factors, which can be learner internal (such as learning processes, individual learning strategies, motivation) or learner external (e.g., social factors such as age, class, ethnic identity as well as factors such as formal language instruction, interaction) (Ellis 1994). Interlanguage is transitional by nature and tends to follow certain patterns of development, certain development sequences and demonstrate great variability in terms of rate of acquisition and outcome. Selinker (1972) suggests that interlanguage is an independent and structured linguistic system containing both errors and non-errors, structures of the foreign language (FL), those that are the result of transfer from the learners' native language (L1) and even structures that do not exist in either the learners' L1 or the FL the learners are learning. The SLA approach to interlanguage, therefore, studies interlanguage as a system of its own and suggests that comparative approaches may preclude researchers' understanding of the systematic nature of the interlanguage process and result in incorrect or misleading assessments; i.e., a comparative fallacy that is revealed by a bias toward the target language (Blej-Vroman 1983) or toward the learners' L1 (Lakshmanan and Selinker 2001).

The advent of computer learner corpus research in the late 1980s caused an upsurge of interest in contrastive and interlanguage studies. Most studies based on learner corpora, however, tend to use contrastive interlanguage analysis, which involves either comparisons between native and non-native data or comparisons of different non-native data; i.e., different learner populations (Granger 2002). Outside SLA, the comparative approach is thus widely regarded as useful since it elicits the specificities of learners' interlanguage. Another emerging method is computer-aided error analysis, which is carried out on learner corpora tagged for standardised error categories. Altenberg (2002), however, warns that interlanguage research should consider all three languages involved - i.e., the learner's interlanguage, the learner's native language and the target foreign language - if it is to offer reliable interpretations of interlanguage features.

Corpus-based contrastive interlanguage studies have managed to enhance our understanding of interlanguage. They have mostly focused on overuse or underuse errors in high frequency vocabulary, spoken and written academic English, modal auxiliaries, link words and phraseology. For example, they have revealed that the unnaturalness of advanced learners' interlanguage, which is so difficult to define, may also be due to the learners' overuse of high frequency vocabulary as well as overuse of a limited number of well formed prefabs (Rigbom 1998; deCock 1998; Petch-Tyson 1998; Cobb 2003). Further, with their use of personal pronouns, learners may signal a different level of personal involvement than native speakers (Rigbom 1998; deCock 1998; Petch-Tyson 1998; Cobb 2003). Contrastive interlanguage analysis has also revealed that overuse of certain coordinators are transfer related when not shared by learners of a different L1 background; others can be teaching-induced by ELT textbooks (Granger 2004).

Corpus-based interlanguage studies have great pedagogical potential. They have informed major learner dictionaries and their effect on teaching materials is also increasing, especially since they can help teachers to decide what lexical, pragmatic and discourse features should be highlighted in teaching and how they should be taught (Flowerdew 2001; Meunier 2002; Nesselhauf 2004). Nesselhauf (2004) further stresses the opportunities teachers have in identifying problem areas with the help of learner corpora and learning about the acquisition processes their students are going through. Both Nesselhauf and Flowerdew also highlight the benefits of data-driven learning that could increase students' awareness of their own problem areas. Since pedagogic materials design is one of the principle tasks of an ESP teacher, the benefits of corpus-based interlanguage analysis could be substantial.

Contrastive learner corpus research is not without its problems. Although learner corpora are not rare (an excellent review is provided by Pravec [2002]), the larger corpora are more readily available only for the more widespread languages. There are still problems with availability of suitable text retrieval software tools; what is more, there is also a serious lack of certain types of corpora: spoken, longitudinal and error tagged corpora, as well as learner corpora covering the field of languages for specific purposes (Granger 2004; Myles 2005). There is also quite a wide agreement that corpus-based studies of interlanguage are often too descriptive and not sufficiently informed by SLA theory (Granger 2004; Nesselhauf 2004; Myles 2005). Furthermore, Nesselhauf (2005) suggests that learner corpus analysis does not allow for comprehension and competence investigations and is more suitable for the investigation of typical learner usage and less for analysing data from individual learners. Cobb (2003) highlights another important problem: the absence of theories that would account for the acquisition processes that intermediate and advanced learners go through. Further compounding the issue is the fact that late acquisition is intertwined with the acquisition of literacy; as a process it is more diverse and the data on native speakers' acquisition of lexis, discourse, and pragmatics is limited.

This paper's aims are thus twofold: to determine whether small ESP learner corpora can assist ESP teachers in assessing the state of their students' interlanguage, and whether the existing resources are sufficient and accessible enough for such an analysis. To this purpose, a small corpus of student essays has been built and analysed by using a contrastive approach. Function words were selected for the analysis due to the fact that these words are members of closed systems, they are frequent, they occur in any text regardless of its topic or field, and they signal relationships between lexical words and larger units of language (Biber et al. 1999). What is more, past research shows that they are useful in approaching large mass of data. Finally, learners tend to meet them quite early in the foreign language acquisition process, therefore examples of functional words as actually used by learners can provide valuable information on the state of their interlanguage.

Although function words belong to a closed class, analysing them all would be beyond the scope of a single paper. Modal auxiliaries were selected for further analysis for two reasons. They are relatively well researched and learners' use of modal auxiliaries have also been documented (Aijmer 2002; Neff et al. 2003; Neff et al. 2004), which would allow for a comparison with the ESP learner corpus data. Relevant research on modal auxiliaries will be reviewed and the corpus of learner essays will be described. Next, based on word frequency data and keyword analysis, the learners' overuse errors will be studied. Finally, relevant implications of the findings will be discussed.

2. Modal auxiliaries and learners

Biber et al. (1999) distinguish central modal auxiliary verbs (*can, could, may, might, shall, should, will, would* and *must*) from marginal auxiliary verbs (*need to, ought to, dare to, used to*) and semi modals (fixed idiomatic phrases with functions similar to those of modals such as *[had] better, have [got] to, be supposed to,* etc.). They also group central modals into pairs to distinguish past time and non-past time (*could, might, should, would* vs. *can, may, shall, will*); the former conveying also the author's stance beside past time. Central and semi modals can be further grouped into three major categories according to their main meaning:

a) permission/possibility/ability: *can, could, may, might*
b) obligation/necessity: *must, should, (had) better, have (got) to, need to, ought to, be supposed to*
c) volition/prediction: *will, would, shall, be going to*.

As regards the distribution of modals, large corpora show that semi-modals are considerably less common than central modal verbs. Both are most frequent in conversation and least in news and academic prose. Nevertheless, studies in academic English have extensively focused on modal auxiliaries as they play an important role in the evaluation and politeness strategies of authors of academic texts. Research has thus confirmed that there is significant variation in the use of modals from discipline to discipline, within research articles themselves, as well as cross-culturally.

Research into learner's use of modals comprises both SLA and contrastive approaches. Bardovi-Harlig's (2005) longitudinal study of learners' use of modality in future expressions follows the SLA approach. Of all means of expressing the future (*will, going to*, present simple in continuous, lexical futures such as *want to, hope to, have to*, etc.) learners used *will* in the written corpus over 2.5 times more frequently than the lexical future. What is more, *will* was used over 4 times more frequently than other future expressions in the oral corpus. She also found great variation in the preferred use of future expressions in individual learners. Lexical futures appear quite early in learners' interlanguage; however, the roles lexical futures play in learners' interlanguage tend to change in time. In early stages they facilitate future expressions, whereas later they bring overt modality to the interlanguage system. The use of *will*, nevertheless, remains the most frequent of all modals expressing the future.

Studies in the contrastive tradition have found that advanced learners tend to overuse modal auxiliaries, yet the values for the individual modal auxiliaries vary according to learners' L1 (Milton and Hyland 1996; Aijmer 2002; Neff et al. 2003; Neff et al. 2004). Chinese learners use *will* and *may* twice as often as native speakers: they use *will* for confident predictions and *may* for denoting possibility (Milton and Hyland 1996). Swedish learners overuse *will, must, have to, should* and *might*; Aijmer (2002) suggests that learners' overuse of *will* might be due to transfer of conversational uses to argumentative genres and that this could be a sign of learners' inability to distinguish informal spoken and formal written forms. She also suggests that learners overuse *must* to sound more persuasive. Neff et al. (2003) noted overuse of *can* by Dutch, French, German, Italian and Spanish learners, with the Italian and Spanish learners showing the highest frequencies. Spanish learners' overuse of *can* was attributed to transfer from learners' native language and the inclusive writer stance, which is typical of Spanish speakers (the latter by Neff et al. 2004).

Attention to ESP learner corpora has been limited to the fields of business English (Connor, Precht & Upton 2002) and academic English (e.g., Flowerdew 1998; Gilquin, Granger & Paquot 2007). To my knowledge, learner corpora in the field of English for tourism purposes are nonexistent, and there is only one learner corpus of materials written by Slovene students: the corpus of Croatian interlanguage by Balažic Bulc (2005), which focuses on learners' use of connectors.

3. Methodology

3.1 The learner corpus

The Tourism Students' Essay (TSE) corpus is an ad hoc non-annotated monolingual learner corpus that was built from essays submitted by 47 first year students of tourism as part of their course requirement; i.e., learner texts that are within reach of every ESP teacher. The essays were written out of class and were, unlike the essays in the International Corpus of Learner English, based on at least three newspaper or magazine articles written in English on a tourism-related topic. Students were encouraged to use dictionaries and a spell-checker. They were also asked to put their ideas in their own words[1]. A suitable text structure was also suggested.

The students were between 19 and 21 years of age and they had all learnt English for 9 years. As the essays were not intended for corpus data, specific data on the usual parameters of learner corpora (opportunities for learning English out of the educational system, mother tongue, other foreign languages spoken, etc.) were not elicited, however, students gave their written consent for including their essays in this corpus during the following academic year. The students' foreign language competence was not tested with a reliable instrument. Based on my teaching experience I would say that most students' English was at levels B1 and B2 of the Common European Framework. However, this assessment should be taken as tentative, especially since students' individual skill levels tend to vary from skill to skill.

The essays are between 1600 and 2000 words long, which exceeds the length of essays included in other learner corpora. As a genre, the essays are not of a clear type: they range from clearly descriptive to more argumentative styles. The essay topic was chosen by the student. Topics such as tourism types, travel trends, groups of travellers, safety issues, mode of travel, etc., prevail.

[1] How far they met this requirement is difficult to determine.

3.2 The target language corpora

For a contrastive analysis, a target language corpus is also required. However, finding a suitable target language corpus turned out to be an impossible task. The Louvain Corpus of Native English Essays (LOCNESS), which is usually used as a reference corpus in such comparisons (e.g., Aijmer 2002; Neff et al. 2003; Neff et al. 2004), is not entirely appropriate: the essays are all argumentative and they cover topics that are not related to tourism, factors which may slightly distort the comparison (Biber 1988; Dagneaux 1995 and Hinkel 1995, both cited in Aijmer 2002).

To compensate for a lack of suitable target language corpus, the TSE corpus was compared to three target language corpora (Table 1):

- the Travel Supplement (TrS) corpus, an ad hoc corpus of travel articles from British and American newspapers that was built specially for this analysis;
- the written component of the British National Corpus (BNC Written);
- the spoken component of the British National Corpus (BNC Spoken).

The comparison with the TrS corpus guaranteed topic correspondence at least to a certain degree and the comparison with the BNC's written component tackled the medium (i.e., written language). The spoken component of the BNC was also used for comparison since past research suggests that students' interlanguage may be closer to the informal registers of spoken language (Aijmer 2002).
The analysis was performed with WordSmith Tools 3.0. using word frequency analysis and key word analysis[2]. Individual words were further investigated with the help of WordSmith's concordancer.

Attributes	**TSE Corpus**	**TrS Corpus**	**BNC - Written**	**BNC - Spoken**
Tokens	82,156	361,059	90,748,880	9,853,249
Types	7,264	28,785	377,384	61,339
Type/Token Ratio	8.84	7.97	0.42	0.62
Text number	47	308	3215	914
Text type	essay	newspaper article	mixed	mixed
Medium	written	written	written	spoken
Average text length	1,748	1,172	28,227	10,780
Text topics	Types of tourism, Tourism impacts, Tourism trends, Safety in tourism, Air travel	Travel, tourism, trends, hotels, holidays, safety, etc.	mixed (representative of general language)	mixed (representative of general language)

Table 1: Corpora attributes

[2] Key words are calculated by comparing frequencies of each word type in one corpus with the frequencies of the same word type in a larger reference corpus. If the difference in a word's frequency in the two corpora is found to be statistically significant, the word will qualify as a key word (Scott 1999).

4. Results

The word frequency analysis of the four corpora showed variation even among the first 10 words (see Table 2 below). The learner corpus exhibited distinct underuse of articles and distinct overuse of *and, that*, and the present forms of the verb *be*. The majority of the first 50 words in written corpora are typically function words. However, of all the three corpora, lexical words were most frequent in the TSE corpus.

	TSE Corpus		**TrS Corpus**		**BNC-Written**		**BNC-Spoken**	
	Word	**%**	**Word**	**%**	**Word**	**%**	**Word**	**%**
1	THE	4.92	THE	5.94	THE	6.38	THE	4.15
2	AND	3.48	AND	2.94	OF	3.24	AND	2.64
3	OF	3.37	A	2.81	TO	2.69	I	2.47
4	TO	2.72	OF	2.44	AND	2.68	TO	2.36
5	IN	2.20	TO	2.24	A	2.24	YOU	2.31
6	IS	1.93	IN	1.80	IN	2.04	A	2.09
7	A	1.76	IS	1.06	THAT	0.99	THAT	1.82
8	THAT	1.46	FOR	0.96	IS	0.98	IT	1.82
9	ARE	1.31	WITH	0.88	FOR	0.92	OF	1.77
10	TOURISM	1.22	ON	0.82	WAS	0.91	IN	1.42
11	FOR	1.10	FROM	0.72	IT	0.85	IS	0.98
12	THEY	0.86	THAT	0.62	ON	0.74	ER	0.90
13	IT	0.83	IT	0.62	WITH	0.70	YEAH	0.82
14	ON	0.80	AT	0.61	AS	0.69	ON	0.82
15	BE	0.71	ARE	0.56	BE	0.67	WE	0.81
16	PEOPLE	0.65	YOU	0.51	HE	0.63	WAS	0.79
17	WITH	0.63	AS	0.52	I	0.57	THEY	0.72
18	HAVE	0.60	I	0.47	BY	0.56	HAVE	0.70
19	WILL	0.59	BUT	0.46	AT	0.54	IT'S	0.69
20	CAN	0.59	WAS	0.45	ARE	0.46	WHAT	0.68
21	AS	0.58	BY	0.44	HIS	0.46	FOR	0.68
22	NOT	0.57	AN	0.38	FROM	0.45	BUT	0.66
23	THIS	0.53	BE	0.36	HAD	0.45	ERM	0.63
24	WE	0.50	HAVE	0.35	THIS	0.45	WELL	0.62
25	THEIR	0.48	HAS	0.32	NOT	0.44	SO	0.62
26	TOURISTS	0.48	OR	0.32	BUT	0.44	BE	0.61
27	ALSO	0.45	THIS	0.31	HAVE	0.43	THIS	0.59
28	MORE	0.43	WE	0.29	YOU	0.42	NO	0.58
29	OR	0.42	ALL	0.27	WHICH	0.39	ONE	0.58
30	BUT	0.41	THERE	0.27	OR	0.38	DO	0.58
31	BECAUSE	0.41	ONE	0.27	AN	0.36	KNOW	0.58
32	BY	0.41	WHICH	0.26	SHE	0.35	HE	0.57
33	WHICH	0.41	UP	0.25	THEY	0.35	THERE	0.56
34	THERE	0.40	HOTEL	0.25	HER	0.34	OH	0.52
35	FROM	0.39	NOT	0.25	WERE	0.32	IF	0.49
36	TRAVEL	0.38	OUT	0.24	ONE	0.28	GOT	0.48
37	ABOUT	0.35	ITS	0.23	THEIR	0.27	NOT	0.48
38	HAS	0.34	CAN	0.22	ALL	0.27	AT	0.48
39	ALL	0.33	MORE	0.22	BEEN	0.27	WITH	0.48

40	I	0.31	THEIR	0.21	HAS	0.27	ARE	0.46
41	OTHER	0.30	WHERE	0.21	THERE	0.26	THAT'S	0.45
42	TOURIST	0.30	THEY	0.20	WILL	0.26	ALL	0.44
43	SOME	0.28	INTO	0.19	WE	0.25	AS	0.44
44	VERY	0.28	IF	0.19	IF	0.24	DON'T	0.42
45	MOST	0.27	MY	0.19	WOULD	0.23	THINK	0.41
46	WORLD	0.27	LIKE	0.18	MORE	0.22	JUST	0.40
47	SO	0.27	TWO	0.17	UP	0.21	YES	0.40
48	SPACE	0.27	IT'S	0.17	SO	0.21	LIKE	0.38
49	ECOTOURISM	0.27	SO	0.17	WHEN	0.21	CAN	0.37
50	AT	0.27	WHO	0.17	WHO	0.20	ABOUT	0.36

Table 2: The 50 most frequent words in the four corpora

Several small corpora were then built comprising travel articles from the TrS corpus of the same size as the TSE corpus to determine whether corpus size pushed lexical words up among the 50 most frequent words. However, there were never more than one or two lexical words among the first 50 words in these small corpora. Therefore, it was concluded that this phenomenon is probably caused by the limited lexical resources of the students.

In her study, Aijmer (2002) elicited only the following modal words: *will, can, would, could, must, have (got) to, should, may, might, ought to* and *shall*. *Have (got) to* was not included in this study since it was impossible to determine whether the author included *has to, has got to* or *had to* in her figures for this modal or not. Table 3 lists the frequencies of the modals in the TSE corpus and the three reference corpora, and Table 4 shows whether the differences between them are statistically significant.

Types of modal	**TSE Corpus**	**TrS Corpus**	**BNC - Written**	**BNC - Spoken**
will	0.59	0.17	0.26	0.20
can	0.59	0.22	0.20	0.37
would	0.13	0.12	0.23	0.28
could	0.14	0.08	0.15	0.16
must	0.14	0.03	0.07	0.06
should	0.15	0.04	0.11	0.11
may	0.06	0.07	0.14	0.06
might	0.02	0.04	0.06	0.08
ought to	0	0	0	0.01
shall	0	0	0.02	0.03

Table 3: Modal auxiliaries across the selected corpora in percentages

Types of modal	Keyness TSE : TrS	Keyness TSE : BNC-Written	Keyness TSE : BNC-Spoken
will	395.1	256.7	403.9
can	249.5	411.3	90.9
would	0	0	0
could	23.2	0	0
must	130.6	43.3	70.4
should	91.3	0	0
may	0	0	0
might	0	0	0
ought to	0	0	0
shall	0	0	0

Table 4: Keyness of the modals in TSE Corpus vs. TrS Corpus, TSE Corpus vs. BNC-Written and TSE Corpus vs. BNC-Spoken (all at p<0.000000)

Learners' overuse of *will, can* and *must* was statistically significant in all three comparisons. However, please note that their keyness - i.e., statistical significance of the overuse - differed. Thus, *will* showed greater keyness in the comparisons with the spoken component of the BNC and the TrS corpus. In other words, students of tourism used this modal far more frequently than native speakers when speaking or the authors of the kind of travel articles, which served as a source for the essays. The reasons for the differences could not be assigned to topical differences or the argumentative features of the essays. *Will* accounts for 2,069.6 occurrences in the academic subcorpus of the BNC Written and 5,915.6 occurrences in the TSE, which is almost triple. *Must* showed a similar picture; however, the values were far lower in all three reference corpora and the greatest difference was observed between the TSE and the TrS corpora. Therefore, it is more probable that the differences in the use of *will* and *must* are due to L1 transfer, students' proficiency levels in the FL or proficiency in their L1 or FL literacy.

The situation was quite different in the case of *can.* The difference was most significant when the TSE corpus was compared with the written component of the BNC while the comparison with the spoken component of the same corpus assigned much lesser significance to this modal's overuse. The discrepancy in the values suggests that it is possible to say that the use of *can* signals a closer resemblance to the typical uses of this modal in spoken language as suggested by Biber et al. (1999) or Aijmer (2002). This, however, does not mean that other factors are not at play. The difference in the frequency of *can* in the TSE corpus and the BNC-Spoken subcorpus is still significant. Therefore, the reasons for the overuse of *can* by students of tourism may be due to L1 transfer, students' FL proficiency levels, their proficiency in L1 or FL literacy or even a lack of familiarity with written genres.

Will, can and *must* are central modal auxiliary verbs (Biber et al. 1999) and actually cover all three categories of meaning expressed by modals: *can* belongs to the permission/possibility/ ability group, *must* to the obligation/necessity group and *will* to the volition/prediction group. Considering that these three modals are also those that learners of English acquire first (at least in the Slovene educational system), their overuse may be due to developmental reasons, which is in line with Bardovi-

Harlig (2005). At FL proficiency levels of B1-B2, learners are still developing their knowledge of modals as well as that of formal registers. Therefore, it seems plausible that other forms of conveying modality have not yet become part of learners' productive language use.

There is another acquisitional factor to be considered: learners' literacy development. It is well known that first year college students' literacy skills are not yet fully developed in their native language. Figueredo (2006) shows that proficiency in literacy skills in the learners' mother tongue can affect spelling in a foreign language. But does the development of literacy skills affect native speakers' use of modals? To find it out, the frequencies of the three modals were looked up in the school essay and university essay components of the BNC.

Type of modal	TSE	BNC - School Essays	BNC - University Essays
will	5915.58	1985.94	1353.27
can	5854.72	2738.01	3219.84
must	1424.12	615.57	505.13

Table 5: Modal verbs per million words in TSE, the school essay and the university essay components of the BNC

The data (Table 5 above) show significant differences in native speakers' use of the three modals[3]. Whereas frequencies of *will* and *must* show a falling trend, frequencies of *can* increase in the university essays. Whether these changes are due to higher literacy levels of native speakers, different topics, differences within the genre of the essay or all of these is impossible to tell as the BNC documentation (Burnard 2000) does not provide sufficient information.

When frequencies of the three modals in the university essay component of the BNC are compared to the Louvain Corpus of Native English Essays (LOCNESS), they show a discrepancy we cannot account for. Nevertheless, we can establish that the data in the essay components of the BNC seem to suggest that proficiency in literacy skills in one's native language may affect the use of the three modal auxiliaries. Therefore, it seems possible that Slovene learners' proficiency levels of literacy in their L1 may affect their use of modals in both Slovene and in English. Based on the present data set, however, it is impossible to determine the nature of this transfer. For that, further research in this area would be required.

The issue of transfer from learners' L1 must be addressed separately from the issue of literacy skills. Is the overuse of these three modals typical only of Slovene learners' interlanguage? To determine this, frequencies of the modals in the TSE corpus were matched against frequencies of modal auxiliaries in the Swedish component of the International Corpus of Learner English (SWICL) and those in LOCNESS (Table 6 below).

[3] The difference between the two components of the BNC was calculated by using log-likelihood and was significant in the case of *must* at the level of $p < 0.001$ and the other two at the level of $p < 0.0001$.

Type of modal	TSE	SWICL	LOCNESS
will	0.59	0.43	0.26
can	0.59	0.38	0.36
would	0.13	0.33	0.24
could	0.14	0.14	0.12
must	0.14	0.13	0.06
should	0.15	0.25	0.10
may	0.06	0.10	0.07
might	0.02	0.13	0.02
ought to	0	0,02	0,01
shall	0 (1)	0,01	0 (2)

Table 6: Modal auxiliaries in TSE, SWICL and LOCNESS (Source: Aijmer 2002)

When the frequencies of *will, can* and *must* in the two learner corpora are compared to the LOCNESS frequency data, a distinct overuse can be observed. *Will* and *can* are overused by Slovene learners of English even more than by Swedish learners. *Must*, on the other hand, shows a similar frequency in both learner corpora. Therefore, the data seem to be in line with Aijmer (2002), who suggests that modals tend to be overused by learners of English but acknowledges that there are differences in the overuse of individual modals as well as in the degrees of overuse.

Finally, the issue of topical and general features should be addressed. If we regard the TrS corpus as a representative corpus of articles on travel, then the comparison with the newspaper component of the BNC could perhaps highlight differences which are due to topic or discipline. The discrepancy between the frequencies of the modals is statistically significant in most cases (Table 7). Especially significant are the lower values of *will, would, could,* and *should* in the TrS corpus. *Can*, on the other hand, is significantly overused in travel articles, which could suggest that the topics students write about in their essays may require them to use this modal more often.

Type of modal	TrS Corpus	BNC - Newspaper	LL[4]
will	1,651	3,922	-953.96
can	2,243	1,597	+109.19
would	1,163	2,394	-434.96
could	767	1,537	-262.35
must	288	458	-39.08
should	429	947	-199.89
may	704	848	-13.38
might	416	381	
ought to	17	22	
shall	6	41	-29.26

Table 7: Modal auxiliaries compared: the TS Corpus and the Newspaper component of the BNC-Written

[4] LL - Log likelihood calculated at http://ucrel.lancs.ac.uk/llwizard.html.

Based on frequency data of modal auxiliaries it is impossible to determine the discourse and pragmatic reasons that lead to the above figures in travel articles. A detailed study would be needed if materials were to be informed by this discrepancy.

5. Conclusions

The first aim of this paper was to establish whether small ESP learner corpora can provide important insights into ESP students' interlanguage. Focusing on the use of modal auxiliaries in the interlanguage of Slovene students of tourism, we have found that students significantly overuse *will, can* and *must*. However, providing a clear-cut explanation for this error turned out to be a difficult task. The overuse of the three modals seems to be a complex error.

First of all, developmental factors may play an important role in the overuse of the three modals. As Bardovi-Harlig (2005) has shown, the use of central modal auxiliaries is more typical of lower proficiency levels in English interlanguage. Since students of tourism tend to overuse one modal auxiliary for each meaning category of modals, it is probable that this overuse error is closely connected with the acquisition of English as a foreign language.

On the other hand, as the comparison of the school and university essay components of the BNC has shown, literacy skills, too, undergo a developmental process that affects the use of modals by native speakers. However, how far the students' literacy skills in Slovene and their literacy skills in English affect their overuse of *will, can* and *must* could not be determined due to a lack of data on literacy acquisition in Slovene as well as a lack of data on modals in Slovene learners' interlanguage at various proficiency levels in English.

However, transfer of literacy skills is only one part of the story. Research has shown that learners' L1 can affect which modals are overused and which not. Italian, Spanish, French, Dutch, German, Chinese and Slovene learners overuse *can* but not the Swedish or the Polish. The use of *will* or *must* shows a different picture across these learner groups. Although we can establish that the differences are probably due to transfer from learners' L1, what exactly in learners' L1 is the cause of this transfer is impossible to determine without detailed analysis of their L1 - as already suggested by Altenberg (2002). A comparable Slovene corpus of essays, however, does not exist.

As regards the text topic, it has been established that it affects the distribution of modal auxiliaries: most of them (*will, would, could, should*) are significantly underused in travel articles with the notable exception of *can*, which seems to be significantly overused in travel articles. Nevertheless, it would be impossible to assign students' overuse of *can* to a topical effect alone as such an effect would need to be accompanied by an underuse of *will*, *must, should* and *could*, which is not the case in the TSE corpus (see Table 4). All in all, the topic of the essays may affect learners' use of the three modals but to determine the nuances of this effect further analysis would be necessary.

Clearly, a definite explanation for the Slovene learners' overuse of *will, can* and *must* cannot be provided. Nevertheless, certain pedagogical implications of these findings can be drawn for this particular group of students. First of all, learners' awareness regarding their overuse of modals should be raised and contrasted with correct native speaker use. Then, lesser used modal auxiliaries and semi modals should be revised or introduced and practised. Such exercises could be especially beneficial because native language transfer seems to be working in combination with language acquisitional processes in this particular case and this may cause persistence in students' overuse errors. Finally, the teaching materials should be assessed and, if necessary, additional remedial exercises should be provided especially for the less frequently used means of conveying modality.

That brings us to the second aim of this paper: the assessment of existing resources that could assist ESP teachers in using learner corpora as a teaching resource. Firstly, building a learner corpus from texts written by students is simple if the texts are submitted in electronic form. However, finding suitable target language corpora is not as straightforward even when the target language is English. Learner corpora tend to be limited to advanced levels of English, written by students of English linguistics and comprising argumentative essays. As such they may not be a perfect match for interlanguage comparisons that could assist in highlighting L1 transfer. Furthermore, free access is allowed only to the Swedish and Polish components of the International Corpus of Learner English.

If we wanted to follow Altenberg's (2002) recommendation on comparing learner corpora with comparable corpora in learners' L1, difficulties would only persist. While FIDA, the Slovene national corpus, for example, is not freely available, FidaPlus is; however, they do not seem to include school or university essays. Furthermore, the online search facility, although helpful for simple word search, does not easily lend itself to queries on frequencies and comparisons across genres as does for example the online search facility of the BYU-BNC (http://corpus.byu.edu/bnc/).

Although ESP for first year students does not contain language that is highly discipline specific, topical specifics of the language still affect both the textual intake of learners and the texts they produce. Therefore, a comparison of the learner corpus with a corpus of relevant discipline specific texts is recommended. Since such corpora are usually not widely available, building such a corpus has to be planned and performed by the ESP teacher. A discipline specific corpus would also be useful for the ESP teacher in the process of material design. Therefore, building a discipline specific corpus is only to be recommended.

Finally, this analysis would have also benefited from non-linguistic data on learners. Collecting these, however, must be planned in advance. More detailed analysis of the phraseology that modals are part of and the discourse features of the student essays would probably produce more information that could assist the ESP teacher in designing teaching materials. Regardless of these limitations and the hurdles for studies like those described in this paper, ESP learner data proved to be informative enough to outweigh the challenges posed by using a corpus approach to interlanguage analysis. Finally, with more ESP learner corpus research and open access (ESP) learner corpora, ESP teaching could more precisely address the needs of the learners and connect these with the specifics of learners' future disciplinary discourses.

References

Aijmer, K. (2002). Modality in advanced Swedish learners’ written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Balažic Bulc, T. (2005). Connectors in students' academic writing in two closely related languages (On the case of Slovene and Croatian language). *Proceedings from the Corpus Linguistics Conference Series*. 1(1), ISSN 1747-9398.

Bardovi-Harlig, K. (2005). The Future of Desire: Lexical Futures and Modality in L2 English Future Expression. In L. Dekydtspotter et al. (Eds.), *Proceedings of the 7th Generative Approaches to Second Language Acquisition Conference (GASLA 2004)*. MA: Cascadilla Proceedings Project.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman - Pearson.

Bley-Vroman, R. (1983). The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity. *Language Learning*, 33 (1), 1-17.

Burnard, L. 2000. Re*ference Guide to the British National corpus (World Edition)*. University of Oxford. Retrieved November 25, 2006, from http://www.natcorp.ox.ac.uk/docs/userManual/.

Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes*, 59(3), 393-424.

deCock, S., Granger, S., Leech, G. & McEnery, T. (1998). An Automated Approach to the Phrasicon of EFL Learners. In S. Granger (Ed.), *Learner English on Computer* (pp. 67-79). London: Longman.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Figueredo, L. (2006). Using the known to chart the unknown: a review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing*. 19(8), 873-905.

Flowerdew, L. (1998). Integrating 'Expert' and 'Interlanguage' Computer Corpora Findings on Causality: Discoveries for Teachers and Students. *English for Specific Purposes,* 17(4), 329-345.

Flowerdew, L. (2002). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam/Philadelphia: John Benjamins.

Gilquin G., Granger S. & Paquot M. (2007). Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes,* 6(4), 319-335.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Granger, S. (2003). The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies? In Granger S., Lerot J. and Petch-Tyson S. (Eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam & Atlanta: Rodopi.

Granger, S. (2004). Computer Learner Corpus Research: Current Status and Future Prospects. In U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics*. Amsterdam - New York: Rodopi.

Han, Z. & Selinker, L. (1999). Error resistance: towards an empirical pedagogy. *Language Teaching Research,* 3 (3), 248-275 .

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Lerot & S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam & Atlanta: Rodopi.

Milton, J.C. P. and Hyland, K. (1996). Assertions in students' academic essays: a comparison of English NS and NNS student writers. *Language analysis, description and pedagogy. Proceedings of international conference organized by Language Centre, HKUST (1996)*. HKUST, 1999 (pp. 47-161). Retrieved November 25, 2006, from http://hdl.handle.net/1783.1/1045.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373-391.

Neff, J., Dafouz, E., Herrera, H., Martinez, F., Rica, J., Diez, M., et al. (2003). Contrasting Learner corpora: the use of modal and reporting verbs in the expression of writer stance. In S. Granger & S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges*. Amsterdam - New York: Rodopi.

Neff, J., Ballesteros, F., Dafouz, E., Martinez, F. and Rica, J. (2004). Formulating Writer Stance: A Contrastive study of EFL Learner Corpora. In U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics*. Amsterdam - New York: Rodopi.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins.

Petch-Tyson S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on Computer* (pp. 67-79). London: Longman.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal*, 26, 81-114. Retrieved November 25, 2006, form http://icame.uib.no/ij26/pravec.pdf.

Rigbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer*. London: Longman.

Scott, M. (1998). *WordSmith Tools*. Version 3.0. Oxford: Oxford University Press.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10 (3), 209-231.

Thompson, P. 2002. Modal Verbs in Academic Writing. In B. Ketteman & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis*. Amsterdam - New York: Rodopi.