CHAPTER NINE

Automatic profiling of learner texts

Sylviane Granger and Paul Rayson

1 Introduction

In this chapter Crystal's (1991) notion of 'profiling', i.e. the identification of the most salient features in a particular person (clinical linguistics) or register (stylistics), is applied to the field of interlanguage studies.[1] Starting from the assumption that every interlanguage is characterized by a 'unique matrix of frequencies of various linguistic forms' (Krzeszowski 1990: 212), we have submitted two similar-sized corpora of native and non-native writing to a lexical frequency software program to uncover some of the distinguishing features of learner writing. The non-native speaker corpus is taken from the International Corpus of Learner English (ICLE) database. It consists of argumentative essay writing by advanced French-speaking learners of English. The control corpus of similar writing is taken from the Louvain Corpus of Native English Essays (LOCNESS) database.[2] Though limited to one specific type of interlanguage, the approach presented here is applicable to any learner variety and demonstrates a potential of automatic profiling for revealing the stylistic characteristics of EFL texts. In the present study, the learner data is shown to display many of the stylistic features of spoken, rather than written, English.

2 Lexical frequency software

The lexical frequency software used for the analysis was developed at Lancaster University (see Rayson and Wilson 1996) as a front-end retrieval system to enable researchers to view semantically (word-sense) tagged corpora and perform statistical tests on frequency profiles produced from those corpora.

The software can provide frequency profiles and concordances (at all levels of annotation) from semantically and part-of-speech (POS) tagged text and has been adapted to display the frequency of lemmas alongside word forms, POS disambiguated word forms and semantically disambiguated word forms. The user can load a file (or set of files) into the program which then displays a frequency profile with relative frequency and a dispersion value (which, in this case, shows how many essays mention each item) (see Table 9.1).

Table 9.1: The word frequency profile

Word	NS frequency	NNS frequency	Overuse or underuse	X' value	Log likelihood	Dispersion
the	*14,912*	*17,728*	*X-*	*29.6*	*29.5*	*702*
Of	*7,645*	*10,282*	*X+*	*17.4*	*17.4*	*702*
to	*7,597*	*9,585*	*X-*	*0.0*	*0.0*	*702*
and	*6,018*	*6,976*	*X-*	*23.7*	*23.7*	*702*
a	*4,726*	*7,034*	*X+*	*76.4*	*77.0*	*702*
i . n	*4,556*	*5,769*	*X+*	*0.0*	*0.0*	*702*
is	*4,465*	*6,518*	*X+*	*55.7*	*56.1*	*700*
that	*3,671*	*4,109*	*X-*	*28.3*	*28.2*	*701*
for	*2,177*	*2,324*	*X-*	*31.8*	*31.7*	*684*
it	*2,116*	*3,270*	*X+*	*52.5*	*53.0*	*693*
be	*2,066*	*2,792*	*X+*	*5.4*	*5.5*	*686*
he	*2,049*	*1,800*	*X-*	*127.6*	*126.6*	*377*
as	*1,978*	*2,368*	*X-*	*3.1*	*3.1*	*674*
not	*1,883*	*2,651*	*X+*	*13.0*	*13.1*	*678*
this	*1,872*	*2,469*	*X+*	*2.0*	*2.0*	*667*
are	*1,682*	*2,701*	*X+*	*60.1*	*60.8*	*668*
they	*1,479*	*2,340*	*X+*	*46.2*	*46.7*	*623*
his	*1,435*	*1,238*	*X-*	*97.7*	*96.8*	*416*
with	*1,396*	*1,614*	*X-*	*5.8*	*5.8*	*663*
by	*1,270*	*1,389*	*X-*	*13.8*	*13.7*	*619*

have	*1,252*	*1,891*	*X+*	*24.2*	*24.4*	*665*
on	*1,228*	*1,702*	*X+*	*6.2*	*6.2*	*658*

Using a classification scheme based on the SGML information encoded in the essay file headers, a user can select subcorpora and hide parts of the text not of interest in a particular study. A typical header is of the type '<p mt=2 tt=1 nr=A1001 >', encoding *mother tongue* (mt), *text type* (tt) and an identification number for each essay. The classification scheme allows the user to display frequencies for different parts of the corpus alongside each other. The X^2 statistic is used to show items whose frequency distribution across the subcorpora is statistically significant. Profiles can be resorted on any of the fields being displayed (including X^2 value). The frequency profile can also be searched for, or limited by, a particular lexical item or tag, for example, to include only lexical verbs by matching on VV.
Values of X^2 are known to be unreliable for items with expected frequency lower than 5 (see Dunning 1993), and possibly result in overestimates for high-frequency words and when comparing a relatively small corpus to a much larger one. In this study the corpora are similar-sized, and results are usually checked using the dispersion value and concordances to take into account the distribution within the corpus. We also use the log-likelihood value (Dunning 1993) which does not suffer the same problems as X^2 does with unbalanced sample sizes.

and sustained". Patients don't	feel	the up and down effect other" stre
the architect mentioned before,	felt	less obsessed with his work and had
ncomfortable in the event that you	feel	you are constantly being viewed as
ing conditions should be like. He	feels	that, "*". How can this be if t
s? Pattullo, on the other hand,	feels	that homosexuals in the militarywo
hat everyone is entitled to. Wall	feels	that, "*". Sexual discriminatio
ech is o.k. does not mean it would	feel	the same way about the amendment
s, especially Liberal Democrats,	feel	that the death penalty is an integr
us crime. Basically, some people	feel	that a strong death penalty through
penalty as immoral, and therefore	feet	that it is unneeded. Although, so
s a dark ring to it. Those who do	feel	that way see pictures of Oliver Twi
y such as Republican Newt Gingrich	feel	that support payments should be sto
this dehumanization factor, many	feel	that orphanages are no place for ch
create lasting relationships, and	feel	a sense of belonging. Speaking of
d to have sex with a class mate to	feel	socially accepted by my peers. My
mate on what our options were. We	felt	the right decision was to get marri
y that a person in a coma does not	feel	pain? Some people have little or n
in life to be breathing, eating,	feeling	, smiling, and most of all loving
would still be life. 1 would not	feel	the same about a terminal illness o
utlook on life. In conclusion, 1	feel	that the restoration of the "Ameri
ne of the users of this system. 1	feel	that this won't help, considering
ofits of the county's transit. 1	feel	that the city might lose more money

Figure 9.1 The KIIC (Key item In Context) display for the lemma 'feel'

To produce a KIIC (Key Item In Context, see Figure 9.1) concordance for an item in the frequency profile, the user simply double clicks on the line in the list. Levels of annotation can be added to or taken away from the concordance lines so that the user can see patterns of tagging, for example, surrounding a key item. Essay headers can also be viewed for each concordance line.

3 Word category profiling

3.1 Word category set

One way of characterizing a language variety is by drawing up a word category profile. This method has been used in previous studies to bring out the distinctive features of learned and scientific English (Johansson 1978, 1985), American vs. British English (Francis and Kucera 1982) and spoken English (Svartvik and Ekedahl 1995).

Claws4, the word category tagging system used for the analysis, employs 134 word category tags,[3] some of which were grouped together for this study, to allow significant patterns to emerge. The reduced tagset contained nine major word categories and 14 subcategories, presented in Table 9.2.

Table 9.2: Reduced word category tag list

N	nouns (common and proper)	
J	adjectives	
I	prepositions	
AT	articles[4]	
D	determiners	
C	conjunctions	
	subcategorized into	coordinating conjunctions
		subordinating conjunctions
P	pronouns	
	subcategorized into	personal pronouns (including possessive and reflexive)
		indefinite pronouns
		wh-pronouns
R	adverbs	
	subcategorized into	prepositional adverbs; particles
		all the other categories of adverbs
V	verbs	
	subcategorized into	lexical verbs (finite forms, *-ing* participles, past participles, infinitives)
		modal auxiliaries *be/have/do*[5]

As appears from the list, five word categories are not subcategorized at all, while the other four have various degrees of secondary coding. Most of the new categories are merged categories. One category, for example, groups all categories of adverbs (general, locative, temporal, etc.) except for prepositional adverbs and particles.

3.2 Frequency of major word categories

Figure 9.2 displays the distribution of the nine major word categories in the native and non-native corpora. Three categories prove to have similar frequencies in the two corpora: articles (AT), adjectives (J) and verbs (V). But the non-native speaker (NNS) writers overused three categories significantly: determiners (D), pronouns (P) and adverbs (R), and also significantly underused three: conjunctions (C), prepositions (I) and nouns (N).[6]

Not unexpectedly, this type of profile raises more questions than it answers. Aside from the question of whether overall similarity of frequency may conceal individual differences, there are questions relating to the over- and underused groups: is it coordination or subordination that accounts for the overall underuse of conjunctions? What types of pronouns are underused? To answer these questions, it is necessary to look both at the grammatical subcategories and the lexical items they contain. This more detailed analysis is the subject of the following section.

4 Significant patterns of over- and underuse

In order to determine significant patterns of over- and underuse, we produced profiles for lemmas in each major word category and subcategory and sorted them in decreasing order of significance. The software also indicates if the lemma is overused by learners (with X+) or underused (X-). Table 9.3 shows the top 20 lemmas in the category of lexical verbs in decreasing order of X^2 value.

The most significant findings resulting from the comparison of word categories and lemmas in the two corpora are summarized in Table 9.4. The table only contains items which are either significantly over- or underused, not those with similar frequencies.

In the following sections these patterns of over- and underuse are interpreted in the light of the results of previous variability studies.

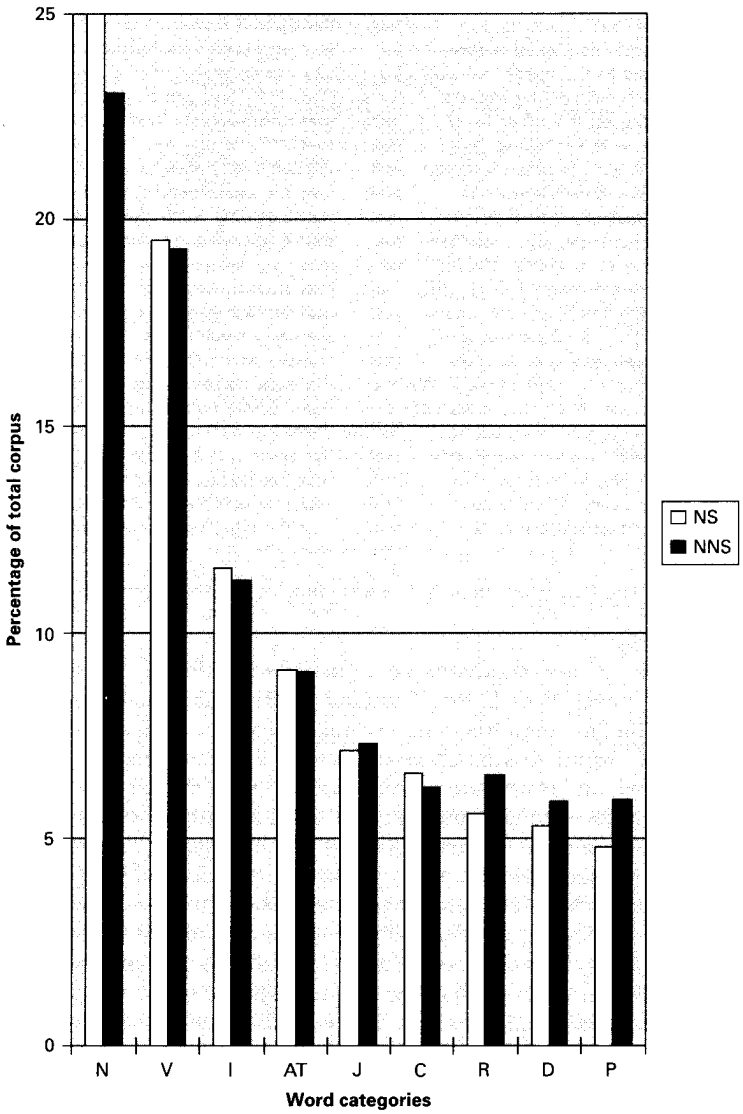

Figure 9.2 Major word category breakdown in NS and NNS corpora

4.1 Articles

In the French learner corpus, the indefinite article *a* is overused and the definite article *the* underused. This proportionally higher use of indefinites by the NNS writers suggests that they are conforming less to the norms of formal writing. In his analysis of word frequencies in the LOB corpus, Johansson (1985: 30) notes that 'category J (learned texts), which has the highest frequency of the definite article, has the lowest frequency of the indefinite article'. These results also demonstrate that an analysis based on major word categories, such as that represented in Figure 9.2, can be very misleading since in the case of articles, it showed no difference between the native and non-native corpus.

4.2 Indefinite determiners and indefinite pronouns

Most indefinite determiners and pronouns are significantly overused by the French learners. A high frequency of such words has been found to be favoured in speech and disfavoured in formal writing. Devito (1966, 1967) notes that speech has more indefinite quantifying words and allness terms, while Johansson (1978: 11, 27) points at the low frequency of indefinite pronouns ending in *-thingl-onel-body* in academic English. Table 9.4 clearly brings out the learners' tendency to opt for the more informal variants of these words: they overuse *a lot* and *lots* but underuse *many*. Similarly, they overuse the indefinite pronouns ending in *-body* but underuse those ending in *-one*, which are more common in writing than the former.[7]

Table 9.3: Top 20 lexical verbs in decreasing order of significance

Lemma	Overuse or underuse	X^2 value	NS frequency	NS relative frequency	NNS frequency	NNS relative frequency	Dispersion
dream	X+	184.2	3	0.00	243	0.08	80
state	X-	112.2	145	0.06	27	0.01	93
think	X+	96.7	261	0.11	666	0.23	418
support	X-	96.0	105	0.05	13	0.00	57
continue	X-	74.3	115	0.05	29	0.01	92
forget	X+	73.9	20	0.01	152	0.05	131
live	X+	72.2	197	0.09	501	0.17	339
speak	X+	66.1	46	0.02	202	0.07	165
imagine	X+	60.8	8	0.00	102	0.04	81
create	X+	58.1	108	0.05	312	0.11	224
believe	X-	55.7	287	0.12	181	0.06	222
argue	X-	53.8	102	0.04	33	0.01	87
realise	X-	51.4	89	0.04	26	0.01	45
allow	X-	41.3	175	0.08	101	0.03	170
disappear	X+	41.3	5	0.00	68	0.02	66
let	X+	40.8	71	0.03	210	0.07	183
run	X-	40.3	66	0.03	18	0.01	54
reach	X+	38.8	39	0.02	144	0.05	124
lower	X-	34.4	32	0.01	2	0.00	10
attempt	X-	33.5	45	0.02	9	0.00	37

4.3 First and second personal pronouns

There is also a very significant overuse in the learner corpus of the first and second personal pronouns. All variability studies associate this feature with the involved nature of speech and point to the low frequency of indices of personal reference in academic writing (see Poole and Field 1976; Chafe 1982; Chafe and Danielewiez 1987; Biber 1988; Petch-Tyson, Chapter 8, this volume and Rayson et al. forthcoming).

Table 9.4: Patterns of over- and underuse in the NNS corpus

	Overuse	Underuse
AT	*a*	*the*
D	**most indefinite determiners**	
	all, some, each, a few, another	*many*
P	**most indefinite pronouns**	
	everybody, nobody, one, oneself, something, everything, a bit, a lot, lots	*no-one, no, anyone, everyone someone*
	first and second personal pronouns	
CC	*but, or*	*and*
CS	**some complex subordinators**	**most subordinators**
	as far as, as soon as, even if	*until, after, before, when, (al)though, while, whilst, whether (or not)*
I		**most prepositions**
	between, towards, without, above, during, of, on, about, before, among in spite of, in front of, thanks to, by means of, till	*for, over, throughout, upon, into, along, out, despite, regarding, per, including, by, off, after, to, amongst, until, up, than*
RP		**most adverbial particles**
RR	**short adverbs of native origin (especially place and time)**	**-ly adverbs**
N		**overall underuse of nouns**
V	**auxiliaries**	*-ing* and *-ed* **participles**
	infinitives	

4.4 Coordination vs. subordination

The general underuse of conjunctions brought out by Figure 9.2 conceals a complex situation. While conjunctions of coordination display both overuse *(but* and *or)* and underuse *(and),* the majority of subordinators; are underused. For reasons which are difficult to explain, the only subordinators that are overused are complex subordinators such as *even if* and *as soon as.* Interpreting these results would require a thorough analysis of each of these conjunctions in context, a task which is beyond the scope of this chapter. However, some results can be interpreted in the light of previous studies. A high frequency of *but* has been found to be a distinguishing feature of spoken language. Chafe (1982) finds over twice as many instances of *but* at the beginnings of idea units in speech as in writing.[8] As stated by Biber (1988: 107) subordination is not a 'functionally unified construct'. Some semantic categories of subordination are strongly associated with speech, and others with writing. It is striking to note that concessive subordinators, which, according to Altenberg (1986: 18) are more prevalent in writing, are significantly underused by learners. It is also noteworthy that the two subordinators which are usually associated with speech, namely *if* and *because,* are not underused by learners, unlike most of the other subordinators.

4.5 Prepositions

The category of prepositions is underused by the learner writers. According to Rayson et al. (forthcoming) use of prepositions differs more than for most other categories between speech and writing. A high proportion of prepositions is associated with the informative and nominal tendency of written language. As appears from Table 9.4, the overall learner underuse hides considerable differences between individual prepositions and again, an in-depth study will be necessary to investigate which prepositions are over- and underused and in what meanings and contexts. Where there are formal-informal doublets, learners again prove to opt for the informal variant: *in spite Of* and *till* are overused, while *despite* and *until* are underused. In addition, complex prepositions, like the complex subordinators, have a tendency to be overused.[9]

4.6 Adverbs

As has now been shown to be the case for many categories, the general overuse of the category of adverbs in Figure 9.2 is the result of over- and underuse of individual adverbs or categories of adverbs. It is mainly short adverbs of native origin *(also, only, so, very, more, even, rather, quite)* which are significantly overused, especially those expressing place and time *(now, ago, always, often, sometimes, already, still, everywhere, here).* The underused adverbs are mainly -ly adverbs: amplifiers *(greatly, truly, widely, readily, highly),* disjuncts (importantly, *traditionally, effectively), modal* adverbs (possibly, *supposedly),* time adverbs in *-ly (newly, currently, previously, ultimately).*

This picture contrasts sharply with the type of adverbs frequently found in academic writing. According to Johansson (1978, 1985), academic writing shows a preference for *-ly* adverbs formed from adjectives of Romance origin which denote concepts other than place and time, and disfavours short adverbs of native origin (especially adverbs of place and time). Learners clearly favour speech-like adverbs. The list of overused adverbs contains eight of the 14 interactional adverbials listed by Stenstrom (1990: 175): *anyway, in fact, of course, indeed, absolutely, really, certainly, now.* It is noteworthy, however, that the underuse of adverbial particles, probably due to an underuse of phrasal verbs, seems to point in the opposite direction since phrasal verbs are typical of speech. A closer look at this category of adverbs is clearly necessary if we are to find out exactly what is happening.

4.7 Nouns

Johansson (1985: 30) contrasts the nominal style of informative prose with the verbal style of imaginative prose. Svartvik and Ekedahl's (1995: 27) study equally links up a lower density of nouns with the category of imaginative texts and conversations. The overall underuse of nouns that characterizes French learner argumentative writing is thus clearly a further sign of a tendency towards oral style. Further research is necessary in particular to assess the rate of nominalizations, which have been shown to figure prominently in academic writing (Chafe and Danielewicz 1987: 99).

A comparison of over- and underused lemmas proves enlightening. Among the underused nouns we find a whole set of items which are normally associated with argumentative writing, such as *argument,*

issue, belief, reasoning, claim, debate, controversy, dispute, support, advocate, supporter, proponent, denial. By contrast, there is overuse of general and/ or vague nouns such as *people, thing, phenomenon, problem, difficulty, reality, humanity* (see Petch-Tyson forthcoming for a discussion of the use of these nouns across several NNS corpora). Such lists clearly hold great potential for ELT materials design.

4.8 Verbs

Though the overall frequency of verbs is similar in learner and native texts, there are considerable differences in the verbal forms used. The first striking feature is the overuse of auxiliaries, a characteristic of conversational English. The second difference concerns lexical verbs, both finite (VVL forms), which are underused and non-finite forms, which display a less uniform pattern, with learners using fewer participle forms, both past participles (VVN) and *-ing* participles (VVG), and more infinitives (VVI) (see Figure 9.3).

This is exactly the opposite of what one would expect in an academic text. Participles are the integrative device *par excellence* (Chafe 1982: 40) and studies such as Chafe and Danielewicz (1987: 101) show that 'language other than academic writing makes considerably less use of participles'.[10] On the other hand, a high frequency of infinitives, which goes together with a high frequency of auxiliaries, is indicative of speech (O'Donnell 1974: 108).

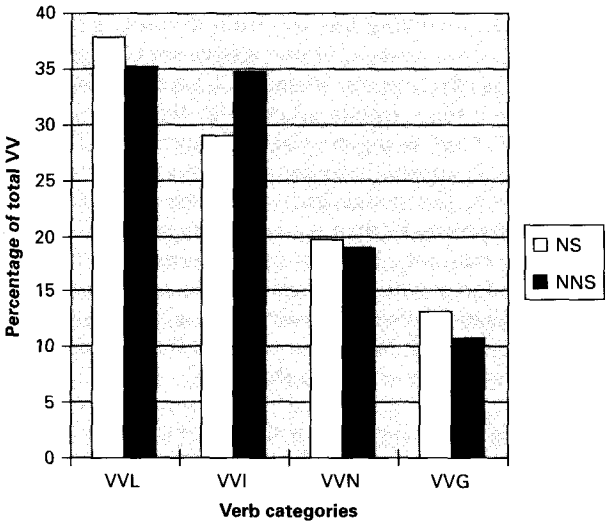

Figure 9.3 Verb forms in NS and NNS corpora

As for lexical variety, a look at Table 9.3 shows that learners underuse some of the typically argumentative verbs - state, *support, believe, argue* - a deficit which contrasts with an unusually high frequency of the ‚cover-all' verb *think.*

5 Conclusion

The automatic profiling technique has highlighted the speech-like nature of learner writing. The essays produced by French learners display practically none of the features typical of academic writing and most of those typical of speech. This conclusion is reinforced by results from other studies, involving learners from different L1s and focusing on other variables (for underuse of the passive see Granger forthcoming a, and for overuse of questions, Virtanen, Chapter 7, this volume).

In our view, two main factors account for this more informal style. On the one hand, there is the possible influence of ELT methodology: the communicative approach to language teaching has put greater emphasis on speech. The models learners are exposed to are more likely to be informal conversations than academic writing. However, this factor alone cannot account for the learners' more spoken style. It merely serves to reinforce a tendency which is essentially developmental. Shimazumi and Berber Sardinha's (1996) investigation of writing by 15-year-old native speakers of English brings out many of the features displayed by the French learners. They conclude that

> The students were asked to produce a written assignment but they ended up producing a piece that has many of the characteristics of spoken language .... they did not show signs of literacy, that is, acquaintance with the formal aspects of written genres.

Orality and involvement are thus more to be viewed as features of novice writing, found in both native and non-native speakers. Whether primarily teaching-induced or developmental, however, the learners' stylistic immaturity has the same remedy, namely greater exposure to good quality expository or argumentative writing, as found, for example, in the editorials of quality newspapers.

Automated quantitative analysis is 'a very accurate quick "way in" for any researchers confronted with large quantities of data with which they are unfamiliar' (Thomas and Wilson 1996: 106). In this article, we have shown that automatic profiling can help researchers form a quick picture of the interlanguage of a given learner population and that it opens up interesting avenues for future research. Do all national interlanguages share the same profile or will each interlanguage have its own? Is the profile constant for a particular national interlanguage or does it evolve across time and if so, how? Automatic profiling applied to a wide range of learner corpora has the potential to help us answer these questions and thereby contribute to a better understanding of learner grammar and lexis.

Acknowledgements

This chapter was written within the framework of the Louvain-Lancaster Academic Collaboration Programme funded by the Fonds National de la Recherche Scientifique, the Commissariat G6n6ral aux Relations Internationales and the British Council.

Notes

1 Crystal (1991: 237) himself suggests extending the concept of profiling to other fields 'to see what might grow'.
2 The non-native speaker corpus consists of c. 280,000 words of formal writing (both argumentative essays on general topics and literature exam papers) by advanced EFL university students of French mother-tongue background. The native speaker corpus consists of c. 230,000 words of similar writing by British and American university students.
3 For a full description of the word tagging system, see Leech et al. (1994).
4 This category includes words which are usually not classified as articles, e.g. *no* and *every*.
5 In CLAWS *belhaveldo* each constitute a class of their own, no distinction being made between their use as lexical verbs or auxiliaries.
6 Throughout this chapter the significance level has been set at 6.63 ($p < 0.01$).
7 A comparison of two subcorpora of the BNC - one representing informal speech, the other informative writing - found there to be a systematic preference for *-body* pronouns over *-one* pronouns in speech and the reverse in writing (except for *nobody* which was found to be more frequent than *no-one* in both speech and writing).
8 The NNS writers' underuse of *and,* also a speech-typical feature, seems to point in the opposite direction. Further analysis of the use of *and* in context will be necessary in order to identify how it is used by the different groups and in what functions it is underused by the NNS writers.
9 One of the reasons why complex subordinators and prepositions are overused may well be that, unlike single word prepositions, they tend to be semantically transparent and have one-to-one equivalents *in* the learners' mother tongue: *by means of = an moyen de; thanks to = grâce à*. Other reasons may play a part as well: the overuse of *as far as* is simply due to the massive overuse of the phrase *as far as X is concerned* by the French learners.
10 A recent study of non-finites in learner writing (see Granger forthcoming b) reveals an underuse of participle clauses in writing by EFL learners from different Lls.