A longitudinal corpus of Swedish university students' written English, some problems and some results*

PHILIP SHAW, UNIVERSITY OF STOCKHOLM

Abstract

To understand how writing skills are acquired requires longitudinal studies, which should focus not only on correctness, but also on style and vocabulary development. In this project I got first-semester students in the English Department of Stockholm University to write early and late in the semester on the same topic. This gave me two samples which should only have differed due to development in the subjects' proficiency. To analyse them I made use of the facilities in the concordancing program Wordsmith Tools (Scott 1996).

The first set of essays contained more different word-types than the second., probably because the writers later on had more uniform ideas of what was expected and perhaps were trying less hard to be original. The code used in the second essays is less personal. This is partly because of changes in the content, but the form of expression is also less personal, with, for example, fewer instances of I think, an index both of less personal writing and of a more written style.

The second essays also conform more closely to the superficial conventions of the academic written code, changes which make the writing more acceptable to an international adult audience. Only two features reflect any deeper development: the small increase in discussion of what text means, and the increasingly selective use of we, which starts to focus on the key function of persuading the audience.

There are a number of problems with the comparability of the samples and with the nature of the essays, but this approach could be combined with more naturalistic longitudinal studies to provide more reliable and valid findings than either could produce on its own.

1 Introduction

When students take a university-level course in a foreign language which they already know quite well, one of the main directions in which their proficiency is expected to develop is towards improved writing skills. Better writing involves a variety of subskills, among which improved correctness, conventional written style, wider and more appropriate vocabulary, effective cohesion, and logical and coherent structure or argumentation are often named. It is relatively easy to assess the state of these skills at a given point, but to understand how they are acquired longitudinal studies are necessary. Thus, for example, Enkvist (1995) examined the development of the spoken and written Spanish of Swedish students of Spanish over two terms and found that their accuracy improved, but that they continued to use a more limited range of grammatical resources.

* A different version of this article has been published in *Studier i modern språkvetenskap*. New Series. Nr 13. Acta Universitatis Stockholmiensis, Stockholm, 2004.

However, the written Spanish in question was short answers to questions and hence always essentially spoken language written down.

Given that Swedish university students of English are often fluent and effective users of the spoken language, development of effective and appropriate writing is a particularly important part of their language training. There have been several longitudinal studies of the development of correctness in their writing. Ruin (1996) examined changes over a nine-week period in the correctness of the English in translations and free compositions. Karlsson (2002) examined similar features in similar texts, but was able to follow her subjects up over three semesters. She found that although there was marked improvement in more salient and learnable features, there was a substantial residue of errors both in, for example, straightforward areas like subject-verb concord, which she ascribed to lack of automatisation, and, to a greater extent, in systems which were complex and hard to make explicit.

Impressionistically, however, third-semester students write much better English than first-semester ones, and if this is not mainly because they are more accurate, it is worth investigating whether the quality of their written language improves in other ways. Only compositions can be used for this. One problem that such investigations can run into is that different tasks and prompts call for different types of language (Peyton et al 1990, Reid 1990), so that if one compares a first-semester and a second-semester essay on different topics, the effects of development are confounded with those of topic-difference.

In previous work Shaw and Liu (1998) attempted to solve this problem by getting subjects to write twice on the same topic. We carried out a longitudinal investigation of the changes in the written English of a cosmopolitan group of overseas students which appeared to take place over a course in academic English. Our subjects, new students at the University of Newcastle, wrote a 200-word ‘essay’ as part of an initial placement test and another as part of a final achievement test some three months later. Both essays had the same prompt. We summed up our results by saying that our subjects’ written language had become more like conventional academic English. Logical connections had become more explicit, the writers’ presence in the text was reduced, and evaluation was more impersonal, but the vocabulary used was not apparently richer or more varied, and the texts were not much longer. The course seemed to have been more effective on the level of style than on that of language development.

The aim of this paper is to extend the longitudinal work of Ruin and Karlsson on Swedish university students' English-language proficiency from error analysis to style and vocabulary development. It adopts a version of Shaw and Liu's method and makes use of the facilities in the concordancing program Wordsmith Tools (Scott 1998).

2 Subjects and Method

In the present project the subjects were first-semester students in the English Department of Stockholm University. The sample was more or less typical of this population and therefore predominantly women, predominantly between 20 and 27 years old, nearly all with at least nine years' study of English, and nearly all with Swedish as L1 or at least the dominant language. A very few of the texts analysed showed traces of other languages than Swedish as substratum, as is presumably typical of Swedish-national students in Sweden today.

All new entrants to the Department take an objective grammar-based 'diagnostic test' and in August 2002 a short voluntary essay question was attached to this. It was clear to students that this was voluntary and for research purposes, but in the circumstances, naturally enough, most wrote the essay. The prompt was:

Please write a short essay (maximum 300 words), in an appropriate style, answering the following question: *Is it true that only rich countries can afford to worry about the environment?*

The prompt is similar to one of those used by Shaw and Liu (1998) and was intended to produce an argumentative text, possibly with some definition of terms.

In December 2002 a retest was administered to students attending a lecture on contrastive grammar. The prompt and the timing were the same as for the initial test. Everyone who took the initial test should theoretically have been present for the second one, particularly as the second test was administered just before a contrastive grammar exam. In fact, however, only about half the possible subjects (some 140) took both tests. A few had missed or not written the initial test, but very many did not attend the lecture in which the second test took place. This may have been because they chose not to attend this particular lecture, knowing the retest would take place, or because they had dropped out of the English course as a whole, or of the comparative grammar lectures, or because of many other factors. In any case the sample tested is more representative of those who complete the first semester course than of those who start it[1].

After December 2002 some 140 pairs of handwritten essays were available. A sample of 60 pairs was word-processed by the author and Mia Lithell (Lithell 2003). The initial essays were combined as one corpus and the retests as another, and the Wordlist, Keywords and KWIC facilities in Wordsmith Tools (Scott 1996) were used to analyse the results.

Wordlist provides a list of all the orthographic words in the corpus analysed. Singular and plural forms, past and present tenses, etc. are all registered separately, so in order to get an idea of the number of different words used in the

[1]Lithell (2003) shows that the initial essays with no corresponding retest were more colloquial and shorter than those with a retest, confirming that the paired essays came from students who were already closer to the norms valued by the Department.

initial and retest corpora the raw wordlists generated by Wordsmith have to be lemmatised. This was carried out manually and crudely. Derived forms were treated as members of different lexemes, inflected forms as members of a single lexeme, so that *effect*, *effects*, and *effected* were members of a single lexeme, despite the likelihood that some tokens were verbal and some nominal, and the probability that some were misspellings of tokens of *affect.* The same principles were applied to both corpora, so one may hope that the result makes it possible to compare the sizes of the vocabularies used in the two corpora, even though counting homographs as single items means that the overall size of each vocabulary is underestimated.

To get an idea of the qualitative differences between the two sets of texts in terms of language code, I made use of the KEYWORDS facility in Wordsmith Tools. This is intended to compare two frequency lists, one for a large reference corpus and one for a smaller test corpus, and it delivers a list of words which are 'key' for the test corpus, that is in a statistical sense significantly more frequent in that corpus than in the reference. In the present case the two corpora were of similar sizes and both rather small. Consequently it is unlikely that the significance statistic in KEYWORDS is valid. However, the method produces lists of words that are considerably more frequent in one corpus than in the other, which is useful heuristically, and these lists can undoubtedly be used to show the *sorts* of words that are more frequent in one corpus than in the other.
Finally, to see what these numerical differences meant in more qualitative terms I used the KWIC facility, which allows one to call up all instances of a character string (word or phrase) and see what its co-text is.

3 Problems

The writing required of the students here is hurried and done without access to reference material or an opportunity for redrafting, which makes it less valid as a sample of their regular writing. However this artificiality is in theory compensated for by the controlled 'experimental' conditions. The two essays should be different only by virtue of changes in the writers' proficiency, since the sample is large enough to compensate for random swings of mood or health and there is little biological maturing over this period at this age.

Nevertheless, there are two probable effects which limit the control of the conditions. The first is that there is a rehearsal effect. Writers have formulated their thoughts on this topic once and it is bound to be easier to reformulate them, thus leaving more surplus mental capacity for attention to the foreign language. I tried to minimise this, as in Shaw and Liu (1998), by not telling the students the topic would be repeated and having the test and retest twelve weeks apart.

The other effect was unexpected. The two essays were written under different conditions, and several students said that they thought the main results would relate to these different conditions rather than any change in proficiency. The

first essay was written under test conditions at the beginning of term, while the second was written during a lecture towards the end of a tiring term. On the second occasion the subjects were not in a test situation or mood and in fact would have preferred a lecture to writing an essay. Many genuinely wanted to know what the test-retest would show about their proficiency development, but some probably did not take the task as seriously as they would have at the beginning of term, and simply did not write as much.

This had an obvious effect on essay length and could have affected other features. The initial essays (hereafter set 1) were longer on average than the retest ones (hereafter set 2). Set 1 (60 essays) contained 13315 words, and set 2 (also 60) 10860, giving mean essay lengths of 222 and 181 words respectively. Length of essay has often been taken as a proxy for the notion 'fluency', (Page 1994, Roy 1992 and the authorities cited there) so this finding is a little depressing. The students' own comments and the considerations mentioned above suggest that writing conditions and student mood are variables that must be taken into account.

4 Results

Set 1 seemed, contrary to our hopes (Enkvist 1996:27), to use a larger vocabulary than Set 2. There were 1337 different lexeme types (as defined above) in 1 and 1155 in 2. This means that the type-token ratio was 9.96 for 1 and 9.41 for 2. One might have hoped that the vocabulary in set 2, if not more diverse, would at least show development, perhaps, for example, containing more nominalisations, but searches for words ending in *–ity, -ation, -ment* (other than *environment*) and *–ance* showed no particular trend.

It is striking, however, that the two sets of essays – on the same topic – actually used substantially different vocabularies. Only 625 lexemes occurred in both corpora, meaning that 712 of the lexemes used in set 1 were used only there, and 530 of those used in set 2 were used only there. Most of these 'unique' words only occurred once in their corpus.

The keywords figures give a clearer picture of the differences between the vocabulary used in the two corpora. Table 1 lists the words that the program identified as markedly more frequent (in occurrences per ten thousand words) in one set than in the other. These are classified by *ad hoc* categories which reveal the differences between the sets.

The table shows that there were three categories in which all forms which reached the criterion were more frequent in 1 than in 2 (abbreviations, proper names, and determiners), and the same effect seems to be visible in the category 'pronouns'. By contrast, the three categories in which there were rather more keywords for 2 than for 1 are lexical verbs, common nouns, and adverbs. Prepositions and adjectives seem evenly matched. There are more auxiliary verbs among the keywords for 2, and *not* was identified as a keyword for set 2,

but this is partly an effect of the lower number of abbreviations: fewer cases of *shouldn't*, other things being equal, mean more of *should* and *not.*

		Keywords for Set 1, i.e. more frequent in 1	Keywords for Set 2, i.e. more frequent in 2
Verbal expressions	abbreviations	6: can't, don't, I'm, it's, I've, shouldn't	0
	auxiliary verbs	1: going	5:cannot, should, be, been, does
	lexical verbs	7, including need, take, and think	13 with none very frequent
Nominal expressions	proper nouns	6: Africa, America, Europe, India, Uganda, USA	0
	pronouns	6: I, my, someone, we, whatever, you	2: everyone, he
	common nouns	14 including, work and people	19 including deal, fact, and means
Others	determiners	an, lot, those	0
	adjectives	7	7 including *poor,* and *rich*
	Adverbs and particles.	3	8 including *not*
	prepositions	2	2

Table 1: Keywords for the two corpora

The three items under 'determiners' -- *an, those,* and *lot* -- illustrate how these data can be used when the numbers are rather low. Set 1 has 46 instances of *an* (34.5 per 10000 words*)* and set 2 has 21 (19.3), and this difference was big enough for the Keywords function to recognise it. However, the corresponding figures for *a* are 225 (168.9 per 10000) and 178 (163.9 per 10000), and all one can conclude is that there seem to be slightly fewer singular indefinite noun phrases in the second essays. For *those,* the sample is small, with 13 in set 1 and 3 in set 2.

Item		Set 1	Set 2
This	Pronoun	28.5	33.1
	Determiner	26.3	38.7
These	Pronoun	0.8	0.0
	Determiner	15.8	21.2
That	Pronoun	23.3	18.4
	Determiner	10.5	9.2
Those	Pronoun	2.3	0.9
	Determiner	7.5	1.8

Table 2: Frequency of demonstratives

Table 2 gives, like the following tables, frequencies per 10,000 words and compares *those* with other demonstratives. It suggests that there is a general shift from *that/those* to *this/these*. Since virtually all uses are anaphoric, this may reflect a shift from the cohesion system of dialogue to that of expository prose, in which *this/these* may predominate.

The same sort of problem applies to *lot (a lot of)*. If it is placed in its system (Table 3) there is a suggestion that colloquial *bit* and *lot* have been replaced by more neutral equivalents.

Quantifiers (large)	Set 1	Set 2	Quantifiers (small)	Set 1	Set 2
A lot/lots	21.8	9.2	A bit	4.5	1.8
much	24.0	27.6	little	3.0	6.4
many	17.3	15.7	few	1.5	1.8

Table 3: Frequency of quantifiers

The keywords data suggest that set 1 essays contained more abbreviated verb forms, more uses of *(a) lot*, more proper names, and more first and second person pronouns, while set 2 contained more of some lexical items. However the changes in lexical items other than proper nouns as a class are, like those under 'demonstratives', generally either rather small changes in frequent items or changes from 0 or 1 in one set to 4 or 5 in the other. However, Table 4 shows that the observations are more robust in respect of the features more common in set 1.

Item	Set 1	Set 2
Geographical proper name	64.6	21.2
It's	31.5	4.6
Don't	39.1	14.7
Doesn't	14.3	7.4
I'm	9.8	0.9
I	98.4	43.3
We	152.5	92.1
you	43.6	21.2

Table 4: Frequency (per 10,000 tokens) of selected items

The personal pronouns *I, we* and *you* are among the most prominent items that changed frequency over the period under investigation. Table 5 lists the lexical verbs which were the heads of verb phrases of which *I* was the subject. In these corpora, *I* is the subject of verbs which either recount personal experience or characteristics (*I have been to Tanzania, I don't know much about this)* or, as is the case for most of the lexical items listed individually in Table 5, mark personal opinions. Expressions like *I think, I believe* are generally markers of 'certainty and authority rather than uncertainty and tentativeness' (Aijmer 2001: 256) and indicate personal commitment. Table 5 thus shows that fairly frequent

types of authorial presence have decreased greatly. *I mean*, which seems to have a metatextual function, is the only phrase that has not decreased proportionately, and this is examined further below.

Head of VP	Set 1	Set 2
think	29.3	12.0
Know	6.0	1.8
Guess	2.3	0.9
Say	9.0	5.5
Believe	11.3	6.4
Mean	3.8	4.6
Feel	2.3	0.9
Be sure/certain	7.5	0.9
Others	27.0	10.1
Total occurrences	131	47

Table 5: Frequency (per 10,000 tokens) of lexical verbs governed by I

In these essays the pronoun *we* is generally used inclusively to mean 'us humans', 'us rich people' or 'us Swedes'. The verb phrases it governs are often modalised and either describe things we can do (ability modals) or things we are exhorted to do (deontic modals). Table 6 shows that non-modalised verbs and verbs with modals other than deontic decreased in proportion to the total decrease in verbs governed by *we*, while verb phrases with *we* and deontic modals stayed roughly the same. Deontic modals after *we* were selectively preserved, as it were, and this may be significant, for Aijmer (2001: 248) says they are 'characteristic of persuasive patterns of argumentation'.

		Set 1	Set 2
Non-modalised		102.9	56.2
With modals		49.6	35.9
deontic	have to	8.3	8.3
	must	0.8	0.9
	should	3.8	5.5
	need to	2.3	0.0
ability	can	22.5	11.0
	could	3.0	0.9
Other	might	0.0	0.9
	will	6.0	1.8
	want to	2.3	6.4

Table 6: Frequency of modal and similar verbs governed by *we*

A similar effect can be observed with *you*, where the number of *you*'s in general is much lower in set 2, but one particular usage appears to have increased slightly. In both corpora a majority of cases of *you* refer to a typical person as an example and could be paraphrased by *one*, as in …. *if you don't have food on your table how much would you care about the environment?* However in set 2

there are 9 cases (out of 23) where *you* refers to an ideal reader as in *If you look on it from the opposite side you will see that the rich countries are the true ones messing with our environment.* Only three out of 55 of the set 1 *you's* seem to exemplify this usage.

The examples with personal pronouns include the orthographic sentence *One man can't save the world we got to work together!* and its naïve energy draws attention to a general change in the usage of the verb *get,* the overall frequency of which has halved from 31.5 to 17.5. Furthermore the uses seem less colloquial: in Set 2 there are no examples of the use of *got to* as deontic modal just exemplified, or of idiomatic uses like *rich countries has got an easier deal,* or *the food they manage to get hold of.*

Use	Set 1	Set 2
Meaning of a word	0	1
The *means* to do something	0	7
I mean 'I wish to say'	3	5
Ambiguous *I mean* 'I wish to say' or Swenglish 'I think'	2	0
Word X means Y	2	2
Text *means* 'implies'	4	13
Total occurrences	11	28
Overall frequency	8.3	25.8

Table 7: Occurrences of *mean*

Among the few lexical items which seemed to have changed frequency enough to be worthy of comments is the lexeme *mean(s).* Table 7 summarises the uses of *mean, means, meaning,* etc. in the two corpora, showing that set 2 had more occurrences of the verb *MEAN* in which the subject was a piece of text rather than the author, and also that the writers in set 2 were beginning to know and use the noun *MEANS.*

Otherwise the lexical items that became more frequent do not seem particularly meaningful because there were small increases in a fairly large number of items which do not obviously have much in common.

Finally, it is worthwhile to look at a set of constructions to see if there were meaningful changes. Cases of *it* as dummy subject standing for a clause with *that* or *to* are worth investigating because they are often considered important as a means for the writer to evaluate propositions. However, an analysis of the forms used (Table 8) suggests only a small increase in the variety of constructions with *that,* and little change otherwise. However, the list of adjectives used only in set 1 – *likely, scary, worrying, certain, obvious, surprising, fair 'relieving', hard, popular* -- and only in set 2 – *cost free, naïve, logical (x2), unreasonable, possible (x2), crucial* -- suggests that the valuation carried out was more likely to be emotionally-based in set 1 and to be rationally-based in set 2.

Constructions	Set 1	Set 2
Copula + (not) true +that	9	8
Copula + other AdjP +that (it is important that….)	10	5
Copula + easy/easier +to	5	3
Copula + AdjP +to (it is important t...)	10	12
VP +that (It has been argued that….)	3	8
VP + to (It costs a lot to….)	6	7
Copula + NP+that (It is my opinion that…)	1	5
Copula + NP+to (It is our responsibility to…)	5	2
Idiom + to ((It is up to us to….)	4	2
Cleft (It is the rich countries that have the means.)	2	5
(ungrammatical)Pro-form (*It is so that*…)	0	1
Total	55	58
Overall frequency	41.3	54.4

Table 8: Occurrences of constructions with dummy *it* as subject

5 Discussion

There were measurable differences between the two sets of essays, and they are interpretable in terms both of what students have learned from their course, and of the conditions of the tests. At the same time a number of questions are raised at least about the significance of essay length and the effects of teaching.

I think there is little doubt that the reason for the retest essays being on average 20% shorter than the initial ones is that the subjects were not in a test situation or mood. This means that essay length in this case was entirely determined by this particular performance, with no implications for proficiency, and in particular none for fluency. Consequently essay length is only a proxy for or measure of written fluency if we can be sure that other things – emotional state, test conditions, etc. – are equal, and that is not easy to be sure of.

The mood of the writers of the retest essays concealed any gains in written fluency that might have been shown by increased length, but it did not conceal considerable changes in the written code. This seems to have become possibly less specific, and clearly less personal and less like spoken language.

The first set of essays contained more different word-types than the second. This could mean that the writers had a wider vocabulary at the beginning of the course than the end, but this is an unreasonable conclusion. It is more likely that the differences between the set of words used by each individual decreased, that is, the essays were somewhat more uniform in vocabulary. This would mean that they had more uniform ideas of what was expected and perhaps were trying less hard to be original.

This is confirmed by the second essays containing fewer concrete examples of specific countries and environmental problems. The reason is partly that they were shorter, but also partly that they focused rather more on general issues and the actual question they were asked.

The code used in the second essays is less personal, partly because of the content. There are fewer accounts of the writers' extra-textual experience (*I have done some travelling for the past three years and two of them I stayed in London*.) and writing process *(I'm sure there is a lot more to say about this topic but I'm running out of ideas.*). But the form of expression is also personal, using for example *as I just wrote* rather than *as noted above.*

Phrases like *I think, I would say, I believe* have been rather closely studied and they seem to be very characteristic of EFL student writing (Petch-Tyson 1998, Andor 2000, Aijmer 2001) as compared with native (or mature, or both?) writing. The decrease in the number of such phrases seems therefore to be an index both of less personal writing and of a more written style.

The second essays also conform more closely to the conventions of the academic written code, by avoiding abbreviations, using *many* for *a lot*, third-person expressions for exemplifying *we* and *you*, and using *got* less. These are all superficial features which teachers of English as a mother tongue drum into school pupils, so the change makes the writing more appropriate and more acceptable to an international adult audience, but hardly reflects any deeper development.

In fact only two of the features that have been investigated do point to better writing in the sense of something that goes deeper or involves more intense meaning-making. One is the small increase in discussion of what text means, or what the writer means by the text, and the other is the increasingly selective use of *we*, which starts to focus on the key function of persuading the audience. One might add to this the use of *you* to draw the reader into argumentation, and possibly decreased reliance on concrete examples.

It is admittedly unlikely that counts of discrete items in corpora could reveal intellectual development. But these data show that we have been quite successful in schooling students to the superficial norms of academic writing, and without losing that focus we should perhaps devote effort to critical development of vocabulary and grammatical variety. Reading these essays one could have the feeling that the subjects had learned to play a part and sound the way they should rather than to adopt values of impartiality and critical thinking.

6 A future possibility

As noted above, there were two confounding factors here. Student mood and test conditions tended to lower performance, and rehearsal factors tended to raise it. In any future replication of this work I will ensure that both essays are done under test conditions, and will attempt to switch topics, so that half the subjects do topic A in test 1 and B in test 2, while the other half do B in 1 and A in 2. This should eliminate rehearsal factors, while neutralising the effect of topic on performance.

7 References

Aijmer, K. 2001. "*I think* as a marker of discourse style in argumentative student writing". In *A wealth of English: Studies in Honour of Göran Kjellmer*, ed. K. Aijmer. Gothenburg Studies in English, 81

Andor, J.. 2000. *Advanced Writing in English as a foreign language.* Ph.D. Diss. Pécs, Hungary: Janus Pannonius University.

Enkvist, I. 1996. "Longitudinell studie över utvecklingen av kongruensböjning hos spanskstudenter på 20- och 40-poängsnivån". In *Språkförståelse. Rapport från ASLA:s höstsymposium, Lund, 9-11 november 1995*, ed G. Håkansson and U. Nettelbladt. Lund: Lund University.

Karlsson, M. 2002. *Progression and Regression: Aspects of Advanced Swedish Students' Competence in English Grammar.* Lund: Almkvist & Wiksell.

Lithell, M. 2003. "Learner language in Swedish student writing in English: a comparative study." Unpublished graduate thesis. Stockholm University.

Page, E. 1994. "New Computer Grading of Student Prose, Using Modern Concepts and Software." *Journal of Experimental Education* 62:2:127-142.

Petch-Tyson, S. 1998. Writer/reader visibility in EFL written discourse. In *Learner English on Computer*, ed. S. Granger, 107-118. London: Longman.

Peyton, J. K., J. Staton, G. Richardson, & W. Wolfram. 1990. "The influence of written task on ESL students' written production.' *Research in the Teaching of English* 24: 142-171.

Reid, J. 1990. "Responding to different topic types: a quantitative analysis from a contrastive rhetoric perspective." In: *Second Language Writing,* ed. B. Kroll, 191-210. Cambridge: Cambridge University Press.

Roy, E. 1992. "Evaluating Placement Exams with a Structured Decision System." *Computers and Composition* 9:2: 71-86.

Ruin, I. 1996. *Grammar and the advanced learner. On learning and teaching a second language.* Uppsala. Uppsala University

Scott, M. 1996. *Wordsmith Tools*. Oxford: Oxford University Press.

Shaw, P. & E. L. Ting-Kun. 1998. "What develops in the development of second-language writing?" *Applied Linguistics* 19: 224-254.