Measuring and assessing the development of foreign language writing competence

PIETER DE HAAN
Department of English, Radboud University Nijmegen, The Netherlands
KEES VAN ESCH
Department of Romance Languages and Cultures, Radboud University Nijmegen, The Netherlands

Received: 8-6-2007 / Accepted: 15-9-2007
ISSN: 1697-7467

ABSTRACT: This paper discusses the development of writing skills of Dutch students of English and Spanish as foreign languages. Essays written in three consecutive years were analyzed for essay length, word length, and type/token ratio — reflecting linguistic competence. A selection of essays was analyzed for argument structure and the use of cohesive devices. These same essays were ranked holistically by experienced lecturers. Students develop linguistic and discourse competences, but differ according to language, proficiency level, and year of study. Assessors' arguments for ranking are related mainly to the students' linguistic competence. The implications of our findings for research and teaching are discussed.

Key words: foreign language writing, discourse competence, writing development, assessment.

RESUMEN: En este artículo se describe el desarrollo de la competencia de escribir de alumnos universitarios de inglés y de español como lenguas extranjeras en los Países Bajos. Ensayos escritos en tres años consecutivos fueron analizados en cuanto al tamaño de los ensayos, el tamaño de las palabras y el type/token ratio, aspectos que reflejan la competencia lingüística. Una selección de estos ensayos fue analizada en cuanto a la competencia discursiva y el uso de mecanismos de cohesión. Los mismos ensayos fueron evaluados y ordenados holístamente por profesores expertos.

Resulta que los alumnos desarrollan sus competencias lingüística y discursiva pero que se diferencian según la lengua que estudian, su nivel de competencia y su año de estudio. Los argumentos de los evaluadores están relacionados sobre todo con el nivel de competencia lingüística de los alumnos. Se discuten las implicaciones de estos resultados para la investigación y la enseñanza.

Palabras clave: escritura en lengua extranjera, competencia discursiva, desarrollo de la escritura, evaluación.

1. INTRODUCTION

Given that the Netherlands have no great tradition in the teaching of formal foreign language (FL) writing in secondary education, students are relatively poorly skilled in FL writing when they enter university. Therefore students still have to learn how to write argumentative essays, in which they must argue specific points. University FL writing courses devote attention mainly to linguistic aspects of writing, such as grammatical and lexical accuracy, spelling, and punctuation. Far less attention is devoted to how texts should be structured into proper arguments.

Regarding their teachers, an additional problem in this context is the assessment of student essays. It is not clear to teachers how they should evaluate the above-mentioned and other aspects of student writing, due to the lack of any clear tradition in the assessment of all relevant aspects of FL writing. FL teachers are language teachers, rather than writing teachers, so they pay almost exclusive attention to linguistic aspects.

It would therefore be relevant to be able to provide these teachers with an instrument for the assessment of their students' essays. As a first step towards the creation of guidelines that help teachers to assess student essays in a more valid and reliable way, in this paper we will compare teachers' holistic assessment of student essays as it is practiced currently in the Netherlands with an analysis of a limited number of linguistic and discourse features (cf. Grant & Ginther, 2000).

Given this situation, the ultimate aim of the research reported in this paper is truly applied in nature as our intention is to study the development of FL writing skill over time by analyzing certain quantitative and qualitative aspects of student essays, and to compare these analyses with holistic assessment by teachers. In the literature review below we first explore the various components and features of FL writing. We then discuss the aspects that have been studied in order to measure development, and the observation periods taken into account. In addition, we discuss one study aimed at comparing linguistic feature analysis to holistic assessment.

Connor and Mbaye (2002) distinguish four competences in writing: grammatical, discourse, sociolinguistic, and strategic competence. Grammatical competence is regarded as the knowledge of grammar, vocabulary, spelling and punctuation. By discourse competence they mean the way the text is structured, especially with reference to how coherence and cohesion are established. Their sociolinguistic competence refers to the appropriateness of the genre, register and tone of the writing. Strategic competence, they feel, is the ability to assess the intended readership, to address them in the appropriate manner, and to present the appropriate arguments. Polio (2001) does not distinguish between these competences, but suggests nine features that are relevant for the assessment of L2 writers' texts, which would seem to be relevant also for the assessment of (the development of) FL writing. These are overall quality, linguistic accuracy, syntactic complexity, lexical features, content, mechanics, coherence and discourse functions, fluency and revision. Some of these, such as syntactic complexity and lexical features, lend themselves to quantitative analysis fairly easily when broken down into more specific categories.

A first attempt at measuring development of FL writing over time was made by Shaw and Liu (1998), who investigated the changes in writing after a two-to-three-month ESL course in a group of L2 English writers from various language backgrounds. They found that there were few changes in syntactic complexity, text organization, lexical variety and the number of errors. Ortega (2003) analyzed 21 studies between groups and 6 longitudinal studies on academic writing in L2 and FL and found that ESL writers tend to produce more syntactically complex writing than EFL writers. In particular, she studied Mean Length per T-unit (MLTU), Mean Length of Clause (MLC), Number of Clauses per T-Unit (C/TU) and the number of Dependent Clauses per Clause (DC/C). She found that, whereas two to three months of university-level instruction may result in small MLTU change in ESL samples and even smaller in EFL samples, it is only by the end of an observation period of nine to twelve months that changes in syntactic complexity may be substantial.

Ortega's (2003) measures had originally been applied by Wolfe-Quintero, Inagaki, and Kim (1998), who studied the correspondence of measures for syntactic complexity with holistic assessments and curriculum-based assessments. Something similar was done by Grant and Ginther (2000), when they studied 90 essays with different scores in the Test of Written English (TWE). These essays were linguistically tagged and analyzed. It was found that the highest TWE scores correlated to linguistic factors indicating a greater linguistic maturity, such as longer essays, greater lexical variety, more cohesion, which appeared from the use of connectors and reference words, like demonstratives, more nominalizations, and a more varied use of tense and mood, more modal verbs, more subordination and passive voices.

Based on the considerations mentioned by Connor and Mbaye (2002) and Polio (2001) and using the same set-up as Grant and Ginther (2000) we carried out several studies into the development of English and Spanish as foreign languages (de Haan & van Esch, 2004, 2005). Using the same prompt as in Grant and Ginther (2000), we studied the relation between teachers' holistic assessment and the occurrence of three text features indicating linguistic competence, in English and Spanish FL student essays (de Haan & van Esch, 2005). Unlike Grant and Ginther (2000), we did not merely study essays with different scores, but essays written by the same students, on the same topic, in three subsequent years.

We found, like Grant and Ginther (2000), that more advanced writers in English and Spanish in general produced more text and longer words. Type/token ratio as a measure for lexical variety turned out to be a not unambiguous measure, which may be related to the fact that Grant and Ginther (2000) and we used different methods of standardizing these scores (see below).

2. Research questions

As we stated above, the ultimate aim of the research reported in this paper is to study the development of FL writing skill over time by analyzing certain quantitative and qualitative aspects of student essays, and to compare these analyses with holistic assessment by teachers.

We have selected three aspects relating to general fluency and lexical maturity, viz. essay length, word length, and type/token ratio, as well as two aspects of discourse competence, viz. argument structure and cohesive devices. This selection is based on the outcome of Grant and Ginther (2000), and Teijeira Rodríguez, van Esch, and de Haan (2005).

Our three research questions, therefore, are:

1) How does FL students' writing develop over time, with respect to essay length, word length, type/token ratio, argument structure, and cohesive devices?
2) What criteria do teachers apply in the holistic assessments of students' essays?
3) What can we learn from the comparison between the abovementioned features found in student essays and teachers' holistic assessment of the same essays?

3. METHOD

The research project of which this study is a part was started in 2002 and is currently envisaged to run until 2008. The first batches of essays were collected in March 2002. All the essays were written by the then first year students (aged 18–19) on the same prompt that was used by Grant and Ginther (2000), asking them to select their preferred source of news and give specific reasons to support their preference. They were allowed 30 minutes to complete this task. In March 2003 the same students (then in their second year) were invited to write another essay on the same topic, under the same conditions, and in March 2004, when they were in their third year, they wrote a third essay on the same topic. The choice of the topic was the consideration that it was broad enough for every student to write about, it does not demand too specialist vocabulary, and it allows students to provide arguments, because they have sufficient general knowledge. The thirty-minute time span was taken over from Grant and Ginther (2000), because it had been shown to be sufficient to write an argumentative text.

The participants in this research were both a group of Dutch-speaking students of English and Dutch-speaking students of Spanish, and a group of university teachers of English and Spanish. The students were studying at Radboud University, Nijmegen, in the Netherlands. For the current study we decided to take into account only those essays that had been written by students who participated on each of the three occasions. For English this amounted to 69 essays (23 students contributed three essays each). For Spanish the numbers were much lower (Spanish is a relatively small department in Nijmegen): only 9 students contributed three essays each, resulting in 27 essays for Spanish.

The choice for the two different FLs was made on the following considerations. In the Netherlands, English is taught at primary and secondary school for a total of eight years, which makes students fairly competent in English when they start their academic studies, so they can be considered intermediate FL learners. Spanish, on the other hand, is not taught at Dutch primary or secondary schools, which means that Dutch students of Spanish start at zero level, so they are beginning learners. Moreover, English and Dutch are very closely related languages. It can therefore be expected that there will be differences between the development of the writing skills of the students of Spanish and that of the students of English. The essays selected for the detailed description above were put before panels of four raters each for the two languages. Both the English and the Spanish panels consisted of two native Dutch teachers, and two native speakers of the FLs in question.

In order to answer our research questions, we studied essay length, word length, type/token ratio, argument structure, and cohesive devices. The analysis of the first three of these was inspired by the abovementioned Grant and Ginther (2000) study. The results of their examination indicated that computerized tagging of linguistic features can be used to reveal detailed differences among proficiency levels. In a similar way we try to measure the development of a number of aspects of students' writing by measuring differences in proficiency level of the same students in consecutive years.

In preparing the data for the answer to our first research question, all the essays were processed with Word and WordSmith Tools to yield basic quantitative data about essay length, word length, and standardized type/token ratio. Standardization is necessary to neutralize differences in essay length. There are various ways of standardizing type/token ratio; WordSmith standardizes by calculating a ratio for every successive sequence of, for instance, 50 words, and then calculating a mean score.

From the total number of essays we selected nine for English and nine for Spanish, for a further inspection of argument structure and cohesive devices. The first batch of essays (March 2002) had been assessed holistically in 2002 by different panels of assessors. On the basis of these assessments students were classified as belonging to any of three proficiency classes: “higher”, “middle” or “lower” (de Haan & van Esch, 2005). Based on this earlier holistic assessment in 2002, we now randomly selected one student from the higher proficiency class, one from the middle class, and one from the lower class, for each language.

We examined the three essays written by each of these students (amounting to nine essays for each language) with respect to argument structure and cohesive devices. The former has to do with the thesis statement, the presentation of arguments to back up the statement, and the presence of a conclusion. By the latter we mean discourse markers and link words (DMLs), pronoun substitution (PS), and recurrence (R) (cf. Teijeira Rodríguez, van Esch, & de Haan, 2005). Cohesive devices organize text and establish relationships between paragraphs and sentences. DMLs are words and phrases like *moreover, in short, on the other hand, por ejemplo* (for instance), *sin embargo* (however), *también* (also), etc. PS refers to the use of words like *it, they, these, eso* (this), *lo* (it), and *las* (them), etc. By R is meant the recurrence of words or expressions either in the form of literal repetitions or by means of synonyms.

In order to answer research questions 2 and 3 (teachers’ assessssment criteria, and the relationship between text features and holistic assessment), the nine essays selected per language were submitted to a panel of four lecturers for blind holistic assessment. Each panel member was asked to assess the essays holistically and rank them. In order for an experiment to possess ecological validity, the methods, materials and setting of the experiment must approximate the real-life situation that is under investigation. Therefore we deliberately chose not to instruct our raters in any way. By not providing them with any kind of check list or rubric we could be certain that each rater would apply his or her own standards, as they would normally do when marking student essays. Holistic assessment would normally imply a simple letter or number rating. We have decided not to use this type of rating because the various essays were taken from different years in the curriculum. A blind assessment of these essays prevented the raters from deciding the level at which these essays were supposed to have been written. Therefore the raters were instructed simply to rank, rather than rate the essays. A mean holistic rank was calculated for the nine English essays and the nine Spanish essays, after which this ranking was compared to essay length, mean word length, and type/token ratio.

4. Results

Research question 1

The answer to our first research question can be found in the results of the quantitative analyses shown in Table 1 (for English) and Table 2 (for Spanish).

Table 1: Three features of general proficiency and linguistic maturity in English foreign language writing in three consecutive years

	Higher (N=7)			Middle (N=9)			Lower (N=7)		
	2002	2003	2004	2002	2003	2004	2002	2003	2004
Essay Length	403.00	426.14	385.14	263.22	378.44	394.44	243.29	374.57	361.29
	76.94	99.15	107.97	68.78	131.65	102.29	76.39	89.95	71.44
Word Length	4.19	4.29	4.36	4.41	4.40	4.41	4.22	4.18	4.26
	0.24	0.15	0.31	0.30	0.26	0.22	0.09	0.18	0.15
Type/token Ratio	75.90	75.69	78.29	77.83	76.24	78.36	77.43	77.01	76.80
	3.24	1.65	4.70	4.74	3.61	2.84	3.20	2.46	4.12

Table 1 shows that all student write longer essays on average in their second year, indicating a progress in fluency, but this trend is continued into the third year only for the middle group. We performed a univariate ANOVA, taking essay length, word length, and type/token ratio as dependent variables and found a significant difference between the higher level group and the two other groups (at $\alpha = .10$) for essay length, but none between the middle level and the lower level group ($p = .889$) We also found highly significant differences between the first year and the second year ($p = .003$) and between the first year and the third year ($p = .008$) , but none between the second and the third year ($p = .733$).

For word length, we only found a significant difference between the middle level and the higher level group ($p = .083$), and between the middle level and the lower level group ($p = .012$). For type/token ratio, no significant differences were found, nor was any significant interaction between year and proficiency level observed.

Table 2: Three features of general proficiency and linguistic maturity in Spanish foreign language writing in three consecutive years

	Higher (N=3)			Middle (N=3)			Lower (N=3)		
	2002	2003	2004	2002	2003	2004	2002	2003	2004
Essay Length	271.00	308.33	323.00	207.00	344.33	382.67	238.00	320.00	410.00
	53.80	98.72	59.62	31.54	38.59	86.25	55.00	87.87	147.89
Word Length	4.63	4.48	4.63	4.39	4.50	4.46	4.47	4.62	4.41
	0.03	0.11	0.03	0.11	0.14	0.17	0.13	0.16	0.07
Type/token Ratio	78.78	78.18	79.00	70.61	75.39	74.58	74.00	77.30	77.09
	1.60	1.99	3.56	2.33	1.76	2.58	3.92	1.64	0.38

Table 2 shows a steady increase in essay length for Spanish on all three proficiency levels, but the development appears to be greatest in the lower group; they have the highest mean score in 2004. Again, we performed a univariate ANOVA on the Spanish data, and we found a statistically significant difference for essay length between the higher level and the middle level group ($p = .071$), and between the first and the second year ($p = .083$), and between the first and the third year ($p = .010$). For word length, we found a significant difference between the higher level and the middle level group ($p = .071$).Unlike the English data, the Spanish data yielded significant differences for type/token ratio between the higher level and the middle level group (*p* = 002.), between the higher level and the lower level group (*p* = .087), and also between the first and the second year (*p* = .090), as well as between the first and the third year (*p* = .099). No significant interaction between year and proficiency level was observed.

In short, the most remarkable difference between the English data and the Spanish data is the absence of any significant findings for type/token ratio in the English data, while these are the most significant findings in the Spanish data. We will return to these findings in the discussion section.

For this description we studied the three essays written by students randomly selected from each of the three proficiency levels established for English and Spanish. These essays were labeled in the same way for either language. Essays A, B and C were written by a lower level student, essays D, E and F by a middle level student, and essays G, H and I were written by a higher level student. Essays A, D, and G were written in 2002; essays B, E, and H in 2003; and essays C, F, and I in 2004. For each language we characterize the three individual students' development.

In order to answer research question 1, how FL students' writing develops over time with respect to argument structure and cohesive devices, we first discuss the argument structures found in the English and Spanish essays, after which we discuss the use of cohesion devices. The lower level English student starts the first essay with a blunt statement of preference for TV; the same happens in the second essay. The third essay starts with a more general introductory sentence about news sources. In all three essays there is a clear summing up of arguments, as well as a conclusion.

The middle level student starts the first essay with a statement of preferred news source, which is TV. This student's second essay starts with an observation about the importance of keeping abreast of current events. The third essay starts with an example of how newspapers cannot be updated regularly, which is a drawback. There is a clear summing up of arguments in all essays; in the second and third essay there is also some emphasis on drawbacks of other news sources. Neither the first nor the second essay ends with a clear conclusion.

The higher level student, finally, starts the first essay with an anecdote about how she used to get up early to read the newspaper when she still lived at home, but that she lacks the time to do this now. This student's second essay start more boldly with the statement of the preferred news source (TV), while in her third essay a preference for radio and TV is stated. In all three essays there are many arguments pro and con. Neither the first nor the second essay ends with a clear conclusion, while the third essay, curiously, ends with the repeated mentioning of the advantages of newspapers.

In the Spanish essays, we see hardly any development of the argument structure in the lower level student's essays from year 1 to year 3. The first essay starts by stating a preference for Internet and then provides three arguments, but fails to draw a clear conclusion. Nor does the poor language help much to make the structure of the essay transparent. The second essay has a clearer argument structure: after an introductory sentence, with a statement about the preferred news source, the student announces the structure of argumentation that follows in the body of the essay. But again, there is no clear conclusion for the text ends quite abruptly with the last sentence about one aspect of television. The same can be said about the third essay, which also ends abruptly without drawing a conclusion. The argument structure of this essay, with a mix of not well-developed arguments in favor of television in comparison to other media, is even weaker than that of the second essay.

The middle level student's essays clearly show a better development of argument structure in the consecutive years. In the introduction of the first essay, the author states a preference for TV. This preference is supported by four arguments, which are stated and elaborated clearly. There is a conclusion, in which the author, oddly enough, introduces other arguments in favor of TV. The second essay starts with an introduction stating the importance of news and TV as a very important medium. Then the author presents seven arguments in favor, and three arguments against TV. A clear conclusion is drawn, although, again, the final sentence is not complete. The overall argument structure of the third essay is fine: the introduction is informative, there are clear arguments pro and con the preferred news source (TV), and it ends with a clear conclusion.

The higher level student's first essay is fairly short, but the arguments are well-structured: the preferred news source (TV) is clearly stated in the introduction, which is followed by only two arguments, after which there is an elaboration into the advantages of the newspaper, and finally there is a concluding sentence, stating the preference for TV once more. This student's second essay, likewise, has a clear structure: it first names Internet as the preferred news source, after which advantages and a possible drawback are mentioned. The essay is concluded well with a clear statement that Internet has everything to offer at a single key stroke. In the introduction of the third essay, the author first states a preference for TV, supported by the argument of speed of information. However, the author then digresses into a discussion of other media. It ends abruptly, without a clear conclusion, which means that there is hardly any development in the structuring of arguments from year 2 to year 3.

As for the use of cohesion devices, we observe the following. In the lower level English student, there is an increasingly frequent use of DMLs in the second and the third essay. PS shows a steady increase from the first through the third essay, while R, especially in the form of literal repetitions, occurs frequently in all three essays.

The middle level English student shows an increased use of DMLs from the second essay onwards; PS occurs frequently in all three essays, while R is mainly achieved by literal repetition. The higher level English student uses a fair number of DMLs in all three essays; PS occurs frequently in all three essays, but, like the other two students, this student uses mainly literal repetitions to achieve R.

The use of DMLs in the essays of the lower level Spanish student shows a varying picture over the years. In the first essay the student uses hardly any DMLs, but in the second essay we find some enumerative phrases and conjunctions . PS is scarce (and sometimes incorrect) over the years, while R is mainly achieved by literal repetitions, which appears to be true even in the third essay, in which the Spanish word for television is repeated literally in almost every sentence.

The essays of the middle level student show a steady increase in the use of DMLs from year 1 to year 3 and the same goes for the use of PS over the years. Also the use of R by means of synonyms instead of literal repetitions is increasing although there are still many literal repetitions in the essay of the third year.

The higher level student's first essay shows a not frequent, but adequate use of these type of cohesion devices. The same goes for the second year essay but in the third essay there is a much more frequent use of DML's. PS is largely present already in the first essay and increases especially from the second essay onwards. There is also a steady increase from year one to year 3 in the use of synonyms to achieve R.

Summing up, the English students show a varying picture of development in argument structure, with the lower level student on the whole presenting the clearest argument structure in all three essays, while the two other students do not manage to present clear conclusions until their third essay. All three English students generally show an increased use of cohesive devices in their writing development, especially DMLs. As such, they can all be said to have developed over the period studied.

The Spanish students, likewise, show a varying picture of development in argument structure and cohesion devices over the three years. The essays of the lower level students show hardly any development, neither in argument structure nor in the use of cohesion devices; the middle level student's essay, however, show a clear development both in argument structure and the use of cohesion devices, while the higher level student shows an ability to structure his arguments clearly in his first essay, and develops this ability especially in the second year. This student's use of cohesion devices shows a steady development from year 1 to year 3. It is perhaps fair to say that only the middle level and the higher level Spanish student show a development in the period studied.

Research question 2

In order to answer research question 2, we present Tables 3 and 4. Table 3 shows the results of the holistic assessment of the nine English essays.

Table 3: Ranking of the essays (1 = most proficient; 9 = least proficient) on the basis of a holistic assessment by a panel of four English lecturers (lecturers I and II are native speakers of English)

Lecturer/ rank	1	2	3	4	5	6	7	8	9
I	F	I	B	C	E	H	D	G	A
II	C	B	E	A	F	G	I	D	H
III	I	F	C	E	G	D	B	H	A
IV	I	C	F	G	H	E	B	D	A
Mean rank:	C	F/I	F/I	E	B	G	H	D	A
	2.5	2.75	2.75	4.5	4.75	5.75	7	7.25	7.75

There are individual differences among the raters, but essay A is deemed the least proficient by three of them. Inter-rater reliability is moderate (Kendall's W = 0.565). The panel's appreciation of the lower level student's work can be called dramatic. Essay A is the least proficient (on the basis of the mean rank), while essay B comes in fifth position, and essay C comes first. Nearly all the raters gave their arguments for their ranking, but their arguments vary considerably, ranging from comments on grammar and vocabulary to comments on paragraph structure and general content.

Table 4 shows the results of the holistic assessment of the nine Spanish essays.

Table 4: Ranking of the essays (1 = most proficient; 9 = least proficient) on the basis of a holistic assessment by a panel of four Spanish lecturers (lecturers I and II are native speakers of Spanish)

Lecturer / rank	1	2	3	4	5	6	7	8	9
I	H	F	I	G	B	D	E	C	A
II	E	G	I	H	B	F	D	C	A
III	H	I	F	E	D	C	G	B	A
IV	E	G	I	B	H	C	F	D	A
Mean rank:	H/I	H/I	E	G	F	B	D	C	A
	2.75	2.75	3.25	3.75	4.5	5.5	6.5	7	9

On the basis of the mean rank it looks as though for each student the second essay is considered better than the first (B is higher than A; E is higher than D; H is higher than G). However, for the middle and lower level students the third essay is thought to be less good than the second one (C lower than B; F lower than E), while for the higher level student the third essay is considered as good as the second one (equal scores for H and I). Inter-rater reliability is substantial (Kendall's $W = 0.625$).

Raters' arguments for their ranking of these nine essays are rather different from each other in number and explicitness. Rater III provides a global judgment of all the essays in terms of coherence and quality of the language, without mentioning any arguments per essay. Rater I makes a distinction between 'planificación' (planning), meaning the presentation and development of ideas in the essay, 'estructura' (structure), i.e. how these ideas are made explicit on paper, 'lengua' (language), and finally 'desarrollo' (development), which this rater uses to refer to the development stage at which the student is. Rater II consistently mentions two criteria: text and argument structure and use of language. Rater IV is the most explicit, stating the considerations that led her to giving a specific essay a specific rank. These considerations refer to a variety of criteria: content and structure of arguments of arguments, the relation between the writer and the intended reader, and use of language, in which aspects such as vocabulary, grammatical correctness and syntactic complexity are explicitly mentioned. For each essay this rater specifies the scores for each of these criteria.

Research question 3

In order to answer research question 3, what can we learn from the comparison of teachers' holistic assessment of the essays with the occurrence of the linguistic features that we studied, we present tables 5 and 6.

Table 5: Quantitative data and holistic assessment of nine essays, written by three EFL students in three consecutive years

	Higher (N=1)			Middle (N=1)			Lower (N=1)		
	2002 "G"	2003 "H"	2004 "I"	2002 "D"	2003 "E"	2004 "F"	2002 "A"	2003 "B"	2004 "C"
Essay length	504	560	549	398	538	448	160	280	341
Word length	4.05	4.16	4.19	4.41	4.23	4.30	4.24	3.90	4.11
Type/token ratio	72.67	76.18	79.40	77.43	69.80	73.50	75.33	76.00	79.33
Mean rank hol. assessment	**6**	**7**	**2**	**8**	**4**	**2**	**9**	**5**	**1**

All three English students wrote considerably longer essays in 2003 than in 2002, but only the lower level student continued this trend in 2004. Word length development was different for each student: the higher level student wrote the shortest words on average in 2002, but used increasingly longer words in 2003 and 204. The middle level student used by far the longest words, but especially so in 2002. In the following years the mean word length was lower. The lower level student, finally, used shorter words in 2003, but longer words again in 2004, though not as long as in 2002. Type/token ratio development was similar for the higher level student and the lower level student, with a steady increase, while the middle level student had a dramatic decrease in 2003, and an improved score in 2004, which was below the score for 2002. All the students were ranked higher in 2004 than in 2002.

Table 6: Quantitative data and holistic assessment of nine essays, written by three SFL students in three consecutive years

	Higher (N=1)			Middle (N=1)			Lower (N=1)		
	2002 "G"	2003 "H"	2004 "I"	2002 "D"	2003 "E"	2004 "F"	2002 "A"	2003 "B"	2004 "C"
Essay length	312	253	337	249	290	404	161	251	287
Word length	4.65	4.52	4.65	4.49	4.69	4.48	4.63	4.39	4.40
Type/token ratio	77.33	78.80	81.00	72.50	75.60	71.75	74.00	79.20	77.60
Mean rank hol. assessment	**4**	**1**	**1**	**7**	**3**	**5**	**9**	**6**	**8**

All three Spanish students wrote longer essays in 2004 than in 2002, but the higher level student wrote a shorter essay in 2003 than in 2002, while for the other two there was a steady increase in essay length. None of the students used longer words on average in 2004 than in 2002, but the development path was different for each; the middle level student used longer words in 2003 than in 2002, while the other two students used shorter words in 2003 than in 2002. Type/token ratio, finally developed differently in each case: the higher level student shows a steady increase, the middle student shows an increase in 2003, but a sharp decrease in 2004, resulting in a lower score than in 2002. The lower level student shows an increase in 2003 and a decrease in 2004, but the score in 2004 is higher than in 2002. All the students were considered better in 2004 than in 2002.

5. DISCUSSION

For the answer to our first research question, how FL students' writing develops over time, we have considered three features of linguistic competence and two aspects of discourse competence. The quantitative analysis of the three linguistic competence features shows that development can be made visible to some extent, but that it is not, at this stage, possible to establish a clear and statistically significant relationship between these features and the students' proficiency level. Although all students wrote significantly longer essays in their third year than in their first year, no interaction could be observed between year and proficiency level. Interestingly, for type/token ratio in the English essays no significant differences were found between either year or proficiency level, while for Spanish we found a significant difference between the higher level and middle level group. The Spanish data did not yield any significant differences for word length, while for English we found longer words in the essays written by the middle level students than in the lower level students.

Regarding argument structure, we can observe development only in the middle level English student, who does not present a clear conclusion until the third essay. The higher level student keeps presenting arguments that do not lead to unambiguous conclusions. The lower level English student, on the contrary, presents well-structured arguments from essay 1. The lower level Spanish student hardly appears to have developed. The middle level Spanish student has developed steadily into the third year, while the higher level Spanish student, who produced a strong text in her first year, has continued to develop in the second year but less in the third year.

With respect to the use of cohesive devices, there is a fairly limited use of DMLs in all the English essays studied, and the only pronoun that is used extensively is *it*, which may have been induced by the specific topic. It is remarkable that in none of the nine essays studied is there any serious attempt at lexical variation; literal repetition is widespread, even more frequent than pronoun substitution. Development of discourse competence would seem to be more visible on the lower level and middle level than on the higher level. The Spanish essays show a gradual increase in the use of DMLs, with the exception of the lower level student, and to a lesser extent, an increase in the use of PS, and a modest increase in the use of synonyms. The latter is especially visible in the higher level student.

Longitudinal studies like ours are hardly mentioned in an overview article of current second and foreign language writing research (Silva and Brice, 2004). Also, Cumming (2001: 4) points out that research into development of individual FL writers is relatively rare. Most research into development is focused on a study of text features in relation to proficiency level. Our study is an attempt at combining the findings of the latter type of research with the former, the assumption being that learners will become more proficient over time. This increase in proficiency, however, is shown not to be self-evident, neither in student cohorts, nor in individual students, as can be inferred from the discussion above.

Our second research question, what criteria teachers apply in the holistic assessments of students' essays, does not have a straightforward answer. All the teachers in our study were shown to focus heavily on linguistic competence, but they varied greatly in the way they commented on discourse competence. This finding confirms our earlier characterization of FL teachers as language teachers, rather than writing teachers.

Our third research question was whether we could establish a relationship between three aspects of linguistic competence found in student essays and teachers' holistic assessment. This has proved not to be the case. At best, there is an occasional correspondence between the holistic ranking of an individual student's essays with a single feature. However, we have seen no consistent correspondence between holistic ranking and all three linguistic features in any of the students.

6. Conclusion

This study has sought an answer to a number of questions relating to the development of FL writing. It is an attempt to shed more light on the development over a number of years of FL writing competence. We have taken Grant and Ginther's (2000) study as a starting point, and have added a longitudinal perspective. Grant and Ginther show that there is a relationship between TWE level and the occurrence of certain linguistic features. On the assumption that students will become more mature and more proficient writers over time, we have sought to establish a relationship between year of study and the occurrence of linguistic features. It has proved impossible to establish this relationship on the basis of the features studied.

A longitudinal study like ours is faced with methodological problems. First of all, there is a need to collect identical data from the same participants over a number of years, which entails the risk of a certain learning effect due to repeated tasks. However, we feel that this risk is outweighed by the advantage of collecting material that will enable us to chart any development precisely. Also, in order to study writing ability, instead of studying one-shot samples it might better to collect data on several different topics, on different occasions in the course of a single year (Hayes, Hatch, & Silk, 2000). However, Ortega (2003) has shown that development is hard to measure over periods shorter than nine to twelve months. In order to make a longitudinal study more reliable it would ideally include more essays, produced by several different cohorts of students.

A problem specifically related to research question 2 is that we do not know exactly how the various aspects of writing contribute to the raters' overall assessment of the essays. This validity problem has also been noted by Polio (2001), who says that it is difficult to make claims about the contribution of each individual aspect of writing skill to the overall text quality. It should be noted that the raters themselves were not instructed, which may account for the variation in the level of explicitness with which they qualified students' work. In spite of this, we found considerable inter-rater reliability, especially in the Spanish panel, so even though raters stress different aspects of student writing, there appears to be an intuitive consensus among them on how to appreciate students' work. This seems especially interesting in view of the debate on the question whether or not to train teachers extensively in order to achieve uniform assessment (Casanave, 2004). It seems to us that experienced teachers are very well able to assess student work without any formal assessing guidelines. However, we feel that it would be beneficial especially to novice teachers to have a reference guide for assessment, based on an explicitation of experienced teachers' considerations in grading student work.

Comparing a holistic assessment with the analysis of a limited number of language features by definition fails to provide a complete picture of writing skills. What needs to be done most urgently, therefore, is increasing the number of aspects included in our analysis, particularly linguistic aspects like syntactic complexity and grammatical accuracy (some raters commented on poor grammar), and lexical features like density and variety. Also the number of discourse features in our analysis would have to be increased (Wolfe-Quintero et al., 1998; Ortega, 2003).

Writing courses might benefit from the results of this type of analysis, as it clearly demonstrates that students do not sufficiently master the use of discourse markers, pronoun substitution, synonyms, and hyponyms. Even in the third year, students are found to use these to a far lesser degree than we think is desirable (and possible). Writing class teachers should include comments on the proper and varied use of cohesion devices in their feedback to written work.

Writing is a complex skill, comprising aspects of language, argumentation, and organization, which will not all develop at the same rate. The complex interaction of these aspects may well account for the great differences in development that we found among the individual students. It is obvious that we need to study more student essays before any clear developmental patterns can be established.

7. References

Casanave, C.P. (2004). *Controversies in second language writing: Dilemmas and decisions in research and instruction*. Ann Arbor: University of Michigan Press.

Connor, U., & Farmer, M. (1990). "Teaching Topical Structure Analysis as a revision strategy", in B. Kroll (Ed.) *Second language writing. Research insights for the classroom*. Cambridge: Cambridge University Press, 26-130.

Connor, U., &. Mbaye, A. (2002). "Discourse approaches to writing assessment", in *Annual Review of Applied Linguistics, 22*: 263–278.

Cumming, A. (2001). "Learning to write in a second language: Two decades of research", in *International Journal of English Studies, 1*(2): 1–24.

de Haan, P., & van Esch, K. (2004). "Towards an instrument for the assessment of the development of writing skills", in U. Connor & Th. Upton (eds.), *Applied corpus linguistics: A multidimensional perspective*. Amsterdam – New York: Rodopi, 267-279.

de Haan, P., & van Esch, K. (2005). "The development of writing in English and Spanish as foreign languages"; in *Assessing Writing, 10*: 100–116.

Grant, L., & Ginther, A. (2000). "Using computer–tagged linguistic features to describe L2 writing differences", in *Journal of Second Language Writing, 9*: 123–145.

Hayes, J.R., Hatch, J.A., & Silk, C.M. (2000). "Does holistic assessment predict writing performance? Estimating the consistency of student performance on holistically scored writing assignments", in *Written Communication, 17*(1): 3–26.

Ortega, L. (2003). “Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college–level L2 writing”, in *Applied Linguistics, 24*: 492–518.

Polio, C. (2001). “Research methodology in L2 writing assessment”, in T. Silva & P. K. Matsuda (eds.), *On second language writing*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 91-115.

Shaw, P., & Liu, E. (1998). “What develops in the development of second–language writing?”, in *Applied Linguistics, 19*: 225–254.

Silva, T. (1993). “Toward an understanding of the distinct nature of L2 writing: The ESL research and its implication”, in *TESOL Quarterly, 27*: 657–677.

Silva, T., & Brice, C. (2004).‘“Research in teaching writing”, in *Annual Review of Applied Linguistics, 24*: 70-106.

Teijeira Rodríguez, M., van Esch, K., & de Haan, P. (2005). “La coherencia y la cohesión en textos escritos por estudiantes holandeses de español como LE”, in *Estudios de Lingüística Aplicada, 41*: 67–100.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Technical Report No. 17. Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.