Academic Vocabulary: A Corpus Analysis Approach

Nahla Nola Bacha
Lebanese American University

Abstract: *The study is a preliminary descriptive exploratory one to analyze a corpus of academic expository and argumentation essays (approximately 1,000 in total) written by students attending the EFL Program at the Lebanese American University in order to investigate the lexis they use. Specifically, the study attempts to describe the most frequent words in the essays in general and according to rhetorical mode. Furthermore, the study identifies frequency counts of 'running words' (content words: tokens) and different words (types) and then standardized ratios of types to tokens. Preliminary results indicate a high percentage of repetition of the same words in the essays irrespective of English course and essay rhetorical mode, with the word 'the' being the most frequent. Also, the type/token ratio is comparable in all analyzed essays indicating the limited presence of lexical variety. Implications for more focus on vocabulary instruction are made for the L2 teaching/learning situation.*

1. Introduction

Few researchers and teachers would deny the importance of vocabulary in students' writing in academic settings. As Biber, Conrad & Reppen (1998:53) state, words are of "central importance …to language learners … That is, learners need to understand how words are actually used, in addition to simple meanings or lists of supposedly 'synonymous' terms." Implications from L1 research for L2 studies indicate that vocabulary is the 'cornerstone of literacy' "…and that instruction has an impact on…vocabulary knowledge…" (Beck & McKeown in Huckin, Haynes & Coady 1995). A further perspective from L1 research is that "…incidental learning from written context may account for a large proportion of vocabulary growth" (Nagy & Herman 1985 in Huckin, et.al 1995). The wealth of research on both L1 and L2 writing (e.g. Connor 1987; Kroll 1990; Grabe & Kaplan 1996; Mukattash 2003) and the writing theories (e.g. Crusius 1989; Rogers 2005) of the expressivists, cognitivists, interactionists, and the social interactionists that have led to the guided, rhetoric, process/product, English for specific purposes methodologies respectively have all influenced the way vocabulary has been taught in the L2 classroom. The analysis of vocabulary in written texts, whether academic or other, is not new, and in fact a wealth of research has been carried out using different approaches upon which textbook writers have developed various teaching/learning methodologies, be they contextual, explicit or implicit (e.g. Huckin et.al. 1995; Lewis 2002; Mukattash 2003; Atari 2004; Malvern et.al. 2004). In this wealth of ongoing research, lexical analyses has not only attempted to describe the type of vocabulary used in different texts, but also relationships with micro-structural elements of texts such as grammar (e.g. Hunston 2000) and with the macro-structure of texts at the discourse level such as cohesion studies over larger stretches of text (e.g. Hoey 1991; McCarthy 1991; Bacha 2002).

There is a growing body of research focusing on student writing, and thus vocabulary in academic settings using corpora methodology in the text analysis (Biber, Conrad & Reppen 1998; Flowerdew 2003; Glendhill 2000; Green et.al. 2000; Lewis 2002; O'Keeffe & Farr 2003; Simpson & Mendis 2003; Ravelli & Ellis 2004 among others). However, Biber et.al. (1998:175) point out the limitation of such studies emphasizing that

> …there are relatively few publicly available corpora of spoken and written texts produced by language learners in natural settings, and even fewer corpora produced by older children and students.

Nevertheless, the need to develop learners' academic lexical repertoire in order to cope with course work at any level has influenced to a great extent the increasing emphasis on research in lexical corpus analysis. The present study hopes to contribute towards this end.

2. Aims and Significance of Study

The aim of this study is to analyze academic expository and argumentation essays written by students attending the four English as Foreign Language Program courses (009, 101, 102, 202) at the Lebanese American University (LAU) in order to describe the lexis used. Specifically, the study attempts to analyze the most frequent words and extent to which repetition occurs in the essays.

This is the first time in Lebanon to the researcher's knowledge that corpus analysis is being undertaken. The importance of this type of analysis is that it can offer teachers an objective measure of their students' lexical repertoire upon which curriculum design and syllabi can be based for the L2 classroom. It thus offers us an understanding of the effectiveness of our teaching/learning situation and possibly hints at some directions we as teachers can take in expanding the lexis repertoire of our students. The author does not profess that the results of the present study will solve any problems in the teaching/learning situation, but it may offer insights which should be treated in the context with objectivity and forethought. Thus, corpora research can help in the classroom in analyzing natural occurring language (e.g. Coniam 1997; Conrad 1999), but as in all linguistic research, used with understanding and a knowledge of its limitations. As Kennedy asserts, "Corpora should be used judiciously for pedagogical purposes, informing instruction rather than determining it so as to avoid the risk of a return to prescriptivism" (1998:290).

3. Background

Corpus or corpora linguistic studies are not new. They have been around for a long time. McEnery & Andrew (1996) give a brief account of historical corpora as way back to the early 19th century when linguists were preoccupied with frequency and comparison studies of various languages and different texts in order to better understand language of different populations and age levels and also to produce dictionaries. However, recent corpora studies have now an added advantage in that larger bodies of text can be dealt with since automation has been introduced. In fact, corpora studies have increased over the last decade and are quickly becoming useful in many aspects of linguistic research (Egbert & Smith 1999; Sinclair 2004).

Basically, corpora research involves analysing larger quantities of text in different genres in an effort to identify differing or common micro and macro structures or the type of vocabulary used in each (e.g. Aijmer & Altenberg 1991; Stubbs 1996; Conrad 2000; Halliday et.al. 2004). Biber (1988) analyzed a wide variety of written (and spoken) texts and argued for six types of surface linguistic variations. For example, the fourth dimension, 'Overt Expression of Persuasion', consists of necessity modals (e.g. must, should), prediction modals (e.g. will, shall), suasive verbs (e.g. agree, arrange, propose..) infinitives (e.g. to change the rule), and markers of conditional subordination (e.g. if..., unless...) which taken all together define a particular text variation. In academic prose, Biber found further variations among the text genres and points out that the linguistic features of the composition rhetorical modes of exposition, argumentation, etc. being different must be taken into consideration in any study (1998:198).

Grabe & Kaplan (1987 in Grabe 1996) replicated Biber's (1988) work on varieties of expository prose to gain an understanding of the text itself and identified four dimensions:1) non-narrative versus narrative context, 2) interactional versus informational orientation, 3) abstract/logical versus situation information and 4) objective versus expressive style. An important finding was that student text types could be identified in ways other than simple counts of individual surface features and then correlating them with writing quality. The study further reported on a text analysis carried out on Freshman final exam essays that the essays were comparable to the Humanities' academic prose on three of Biber's (1988) dimensions and thus in part were following genre expectations of the academic community (Grabe 1996:48).

McCarthy (1991:148) mentions the usefulness of corpora information in text studies and states that

> ...the existence of huge computerized corpora of written material such as the over 200 million word Birmingham Collection of English Text (the basis of the Collins COBUILD dictionary project) and corpus-building over the years has led to an interest in detailed taxonomies of textual types.

Sinclair (1994, 2004) argues for a re-examination of research methods to exploit the large amount of data now available and to focus more on larger stretches of discourse in the various fields reexamining traditional classifications. Francis (1994), using Cobuild corpora data, shows how a cohesive device called labelling common in the press and argument connects the discourse across clause boundaries by retrospective labels (pointing backward, e.g. *this problem*) and advance labelling (pointing forward, e.g. *three reasons*). Labelling, it is argued has an important organizational function which may differ according to the genre. The significance of this type of research is that text structure is viewed as related to and influenced by genre.

Recently, corpora studies have concentrated on lexicogrammatical issues (e.g. Halliday et.al. 2004) especially in translation studies (e.g. Bowker 1999) and also in finding relations between the use of lexis and grammar (e.g. Hunston & Francis 2000), the results of which text book writers have exploited for the improvement of academic writing. Specifically, with the availability of large corpora, lexicography studies have been able to concentrate on quantitative studies of frequency counts of lexis in texts of different genres and at different historical periods (McEnery & Wilson 1997; Biber et.al. 1998). Comparisons among different texts not possible in the past are done easily as is studying the company certain words keep or collocates with as well as morphological studies providing a wealth of information for textbook writers and teachers (Boguraev & Pustejovsky 1996; McEnery & Wilson 1997; Biber et. al. 1998).

Thus the corpus or corpora terms and what it stands for has become an important part of textual linguistic analysis which can help in the teaching/learning situation (Conrad 1999). McEnery & Andrew (1996) define a corpus "as any body of text", but go on to qualify that when the term 'corpus' is used in the context of modern linguistics it takes on four main aspects: sampling and representativness (also discussed by Biber 1993), finite size, machine-readable form and a standard reference. Also, corpora can be unannotated (in plain text) or annotated for different types of linguistic form that could be under study and quite lengthy and detailed annotations have been described (Garside, Leech & McEnery 1997). Quite popular in recent lexical corpora studies has been the focus on lemmatisation permitting the researcher to examine the variants of the same word; e.g., the lemma *go* has the morphological irregular forms of *went, goes, going, gone.* Researchers must consider in their frequency counts whether they want each of these forms to be considered as separate words or all as one word in a corpus. Parsing and tagging are two other interesting aspects of corpora research, the former dealing with tree diagramming sentences into their different parts into noun phrases, verb phrases, parts of speech and so forth, while the latter deals with annotating the prefixes and suffixes of words which allows the researcher to examine different parts of speech in text.

Of course genre and register do affect any type of corpus analysis (e.g. Stubbs 1996; Biber et.al. 1998; Conrad 2000) and thus any analysis must consider the genre under question or make qualifications in any conclusions. Biber et.al. (1998) mention that Grabe & Reppen (in Biber et.al 1998) have made a corpus of over 5,000 non-native elementary student written essays obtained from forty classes in fifteen towns in Arizona, USA. The study shows how these students wrote on different topics in expository and argumentation rhetorical modes and then were analyzed. Some results indicated that the length of the student essays in

T-units (number of words) averaged 9.6 in third grade and 10.8 in sixth grade with a total average number of words of 76.8 and 113.5 respectively. An important finding in these children's texts is that there was a high frequency of *and* in initial position of sentences giving an indication that coordination is prevalent in younger students' writing. Also, the researchers mention that the writing was very similar to the oral mode, lexical density and sophistication being absent which is the case of L1 Arabic non-native speakers of English in the Lebanese context.

Teachers also find vocabulary important although we know the ongoing complaints as to the 'limited' repertoire many of our students are showing in their writing at any level. However, students seem to have a higher perception of their vocabulary abilities in the context of the essay (Johns, 1981; Zughol & Hussain 1985; Sa'Addedin, 1989; Grabe & Kaplan 1996).

3.1. The Lebanese American University context

Some research done at LAU is indicative of how students rate their vocabulary level, learn vocabulary and perceive its importance in an academic context. A survey was carried out in the English as Foreign Language (EFL) Program at the Lebanese American University during the Fall Semester 1999-2000 to find out the perceptions of both faculty and students of students' writing ability in both the English courses and in other courses in the university along a range of variables including vocabulary. Results indicated that faculty teaching in the disciplines as well as in the English courses (N=48) rated their students' vocabulary abilities as significantly lower (p=.001) on both the use of correct vocabulary and varied vocabulary (synonyms/antonyms etc.) in academic writing than students did (stratified random sample N=980) using the Mann-Whitney statistical test. (Bacha 2000b).

Another study focused on how students learn lexis or vocabulary and how they relate this to academic literacy specifically to their own essay writing experience. Towards this end, a survey was carried out on a stratified random sample of 155 students (Bacha 2001) attending the basic English course 009 in the EFL Program at the Lebanese American University. It must be noted that the survey was an open-ended one, thus not restricting the answers and thus students would indicate their own responses which were later tallied. The following results are reported according to the four survey questions.

1) On how they (students) learn vocabulary, *reading* was the most statistically significant (p=.01) when compared to the *dictionary, media, speaking, listening, other university courses, memorization, writing, tests* and *French* using the Friedman's statistical test for related samples. This result is surprising since it is a common fact that our students do not 'like' to read and often do not read their set home assignments. Probably, this result is a consequence of the teachers often telling their students that one's vocabulary repertoire widens through reading exposure. It is not unusual for students to look up words constantly in the dictionary thinking that this would increase their

vocabulary repertoire. It is also common that our students at LAU are of the opinion that French has a negative influence on writing in English, and thus on their use of vocabulary. However, studies have indicated that students who have studied in the medium of French in their high school studies and have shifted to English at university show no significant difference in use of correct and varied vocabulary in their essays when compared to those that did their high school studies predominately in English, and in fact there could be positive transfer (e.g. Odlin 1989; Bacha 2002).

2) On why they (students) thought vocabulary was important in academic writing, students significantly (p=.01) indicated, *expression*, when compared to *university courses, richness, clarity, knowledge, interesting, high grades, less repetition, career, logic* using the Friedman's statistical test for related samples. Students place importance on good expression of ideas, and the importance of choice of appropriate words to this effect probably due to teachers emphasizing this for both oral and written communication.

3) On what they believe their teachers expected in 'good' essay writing, although they indicated *good organization* as the most significant using the Freidman's statistical test (p=.01), *vocabulary* took priority over *logical ideas, correct grammar, support, sentence structure, mechanics, less repetition*, but the result was not significant. Probably, organization is emphasized in the English classes.

4) On how they viewed vocabulary was related to academic literacy, 70% of the students responded that they *did not know,* while the rest related it to *expression of ideas, logic, technical language, culture and knowledge in society.* Perhaps, the students were not aware what *academic literacy* means, an indication of their limited lexical repertoire.

All in all, although students perceive vocabulary as being important in their writing as indicated in the foregoing study, many are unaware of how vocabulary can best be learned or the degree of its centrality in academic literacy.

4. Problem and Research Questions

The main problem that is faced by L1 Arabic non-native speakers of English and students at the Lebanese American University is no different, a need to expand their lexical repertoire. Both English and discipline faculty complain of the limited lexical repertoire of their students and that means should be found to help widen the range of their students' vocabulary (e.g. Zughol & Hussain 1985; Bacha 2000a; Bacha 2002). The present study will focus on analyzing student essays to substantiate the degree of limitation of the lexical repertoire according to two main classes of words: content and function words, e.g. of the former *school, problem* etc. and the latter *the, of, that* etc.

Research does show that there are two types of vocabulary that any individual can make use of: passive and active. Students may comprehend well what they read (passive vocabulary knowledge), but how well do they actually use (active knowledge) the words they 'know' and to what extent do they use content words? These remain concerns for many English programs especially in an academic context. Two main research questions are raised in the present study.

1. What are the most frequent words in the student essays under study?
2. How much repetition of the same word is there in the student essays under study?

These research questions are in line with Biber's et.al (1998) second of the six posited research questions that they mention as part of recent corpus-based research. Biber et.al (1998:23) phrase the question as "What is the frequency of a word relative to other related words?" and go on to say the information obtained "…can be especially useful in designing teaching materials for language students".

5. Methodology

The present study is a description of "…naturally occurring phenomena without experimental manipulation" (Seliger and Shohamy 1989) but investigating two specific areas: frequency and repetition. It might be argued that preliminary descriptions using computer tools is a simplistic and not valid research method. However, there are researchers that do argue for the effectiveness of corpora based research for the classroom (e.g. Conrad 1999; McCarthy & Carter 2001). Stubbs (1996:232) argues for the importance of descriptive corpora studies in linguistic inquiry and states

> "…even preliminary investigation and description are useful. When computer methods are used to study large corpora, they may confirm what was suspected or known all along…By transforming the data, they can generate insight".

And further states that along with confirmation and insight, the description "...will usually provide vastly more detailed information than would otherwise be possible".

Biber et.al (1998) also give the characteristics of corpus-based analyses which the present study follows. They list four important analytic features: 1) patterns in natural texts, 2) large amount of text referred to as a "corpus", 3) use of computers, and 4) quantitative and qualitative techniques. In line with these analytic features, the natural texts in this study were the essays that the students wrote in class; a large corpus was collected, the computer was used and although the analysis was mainly quantitative, some although limited qualitative analysis was done in actually reading some of the essays and commenting on their lexical sophistication as identified by range in an academic context. For example, the word *good* is often used and repeated in students' writing. To widen the range and/or sophistication of students' writing, synonyms such as *appropriate*, *suitable* etc. could be used depending upon the context; however, students find it difficult to use a wider lexical repertoire.

5.1. Context of the study

The EFL English Program at the Lebanese American University (with three campuses: Beirut, Byblos and Sidon) offers three composition courses at the Freshman level (009-remedial- Paper-based TOEFL score 524-574 –computer based score 193-230), 101 - TOEFL score 575-624- or 233-260) and 102 – TOEFL score 625-674 – or 263-297) and one at the Sophomore level 202 (TOEFL score above 675 or above 297) in which students learn general academic English to cope with their university courses. The main tasks include writing of paragraphs and essays in the expository mode in the more basic English courses, a research paper in English 102 and argumentative and critique essays in English 202. The required texts used in each of the English courses are English 009 (remedial English), Langan 1997; English 101, McWhorter 2003; English 102, Buscemi & Smith 2002; English 202, Behrens & Rosen 2004).

Although research in English for Specific Purposes, specifically English for Academic Purposes views the purpose of such similar programs to prepare students for both the academic and outside professional communities in the language skills (Jordan 1997), the LAU EFL Program focuses on teaching/learning general academic English.

5.2. Participants

The students who wrote the essays were attending the English courses mentioned and placed into the English courses according to their scores on the LAU English Entrance Exam. Thus, some of the students do not necessarily have to take all four courses. Students in any one English class are enrolled in different majors in the university in four Schools: Arts and Sciences, Business, Pharmacy and Engineering and Architecture. The majority of the students are L1 Arabic non-native speakers of English with some having followed high school studies mainly in the medium of French (referred to as French educated) and others in English (referred to as English educated). A few have done their high school in both mediums of French and English (bilingual). However, the high school study language, as it is referred to, was not one of the variables in the present study. However, in ongoing research, it could be an interesting variable to consider.

5.3. Procedure

Beginning the Fall 2001 Semester (October) and ending in the Spring 2002 (May) a total of 1,158 essays (Sample 1) from the various sections of the English courses at the university were typed into the computer into Ascii files. These essays had been written in one sitting of one hour each. Each essay was coded for English course and beginning or end of each of the two semesters (each semester being a four month period). Only the essays from English 202 were argumentative; all others expository (see figures below for numbers). Expository essays included rhetorical modes of narration (topics such as *narrating an incident or event*), description (*describing a person or place*), cause and effect (*reasons and results of loneliness or choosing a major*), and comparison and contrast (*comparing and contrasting two cultures or universities*). Argumentative essay topics included *arguing for or against 1) whether justice exists, 2) schools should educate or train for the workplace, 3) obeying authority, and 4) censorship of the press.*

Since topic and rhetorical mode affect results, the researcher was able to obtain two additional samples (N= 22 each giving a total of 44) from one of the English 101 classes in different rhetorical modes, illustration (topic: *illustrate how man tries to overcome problems in daily life*) and cause-effect, (topic: *reasons and results of war*) written by the same students at the end of the Fall 2001-2002 Fall Semester (Sample 2). Also, at the same period, it was possible to obtain a third sample from two of the English 009 classes, consisting of 36 essays at the beginning of a semester and another set of 36 essays from the same classes and students at the end of the semester (Sample 3); the topic (*Give the causes and effect of disagreements between teenagers and their parents*) and the rhetorical mode (cause-effect).

The total number of essays obtained for the study was representative of any one semester and any one set of students attending the various English courses by study language, major and gender although these variables were not included in the present analysis. A fourth sample, two scanned passages from each of the respective English course textbooks were also used to compare the student essay writing to 'models' of writing and thus level and use of vocabulary for each of the English courses (Sample 4).

Each student essay was computer grammar and spell-checked so that there would be no negative influence from the latter in the analysis and the essays saved as a text file. Textbook passages were scanned and saved as text files.

5.4. Data analysis

Specifically, in corpus terms, the data were first analyzed for percentage of repeated words. The researcher focused on frequency counts of 'running words' (content words, referred to as tokens) and different words (referred to as types) and then ratios of tokens to types analyzed. For example, Scott (1998:19) calculates that

> "If a text is 1,000 words long, it is said to have 1, 000 tokens. But a lot of these words will be repeated, and there may be only say 400 different words in the text. Types therefore are the different words. The ratio between types and tokens in this example would be 40%".

As ratios naturally vary with the length of the text or a corpus being examined, and in order to obtain meaningful information, a standardized type/token ratio is computed every **n** words and an average standardized type/token ratio computed. To obtain the standardized type/token ratio, counts were normed to a basis of 500 words since the average length of the texts was computed and found to vary between 475 and 540 across the English courses.

Biber et.al. (1998:264) state that "…frequency counts should be normed to the typical text length in a corpus".

Second, to find out the most frequent words, word lists by percent frequency were made for the essays. The default in the Wordsmith Tools was kept at 1 minimum and the maximum number (2million plus) which would include all the words in the texts. All the variants of each word were considered as different words; that is, the analysis did not consider lemmatization, and hyphenated words were considered as one word. Since these are learner texts, possible errors made might have influenced the results.

6. Results and Discussion

Results are given based on total essay samples obtained from each of the English courses except where indicated. Since results showed no differences between beginning and end of semester essays, these were not reported on in any detail.

6.1. Word frequencies (research question 1)

Table 1 indicates the ten most frequent words when computed as a total in each of the English courses and the percent frequency of each of the words (n=1 represents the most frequent word) with the word *the* being the most frequent across the English courses. The scanned texts were comparable. (The capitalization of the words in the tables does not necessarily mean that the words begin sentences.)

Table 1: Most Frequent Word in Total Sample Essay Corpus by English Course

N	**009** **N=634**	**%**	**101** **N=150**	**%**	**102** **N=69**	**%**	**202** **N=264**	**%**
1	The	4.52	The	5.59	The	6.61	The	5.82
2	To	3.35	And	3.34	And	3.16	To	3.09
3	And	3.05	Of	3.09	Of	2.91	And	2.87
4	In	2.34	To	2.77	To	2.86	Of	2.80
5	A	2.28	A	2.49	In	2.33	In	2.47
6	Of	1.91	In	2.09	A	2.12	Is	2.35
7	That	1.65	Is	2.05	Is	1.81	A	2.16
8	I	1.64	That	1.70	That	1.24	That	1.85
9	Is	1.60	Are	1.20	Are	0.95	It	1.07
10	For	0.93	For	0.99	Be	0.85	Be	1.00

It seems that in general academic texts, the function words are the most predominate in both student essays and the scanned passages taken from the required English course textbooks. However, the scanned book passages did show that the content words began to appear more frequently at about n=20 and the student essays across the English courses at about n=50 indicating that the latter had fewer content words. It was interesting to note that in the English 101 essays in the rhetorical modes of *illustration* and *cause effect* the content word *problems* appeared as the ninth most frequent and *war* the fifth respectively. Although this finding is based on a limited sample in one genre, be it expository or argumentation, the results indicated that when only one essay topic is examined a 'key' word appears higher in frequency than if a whole corpus is taken including a variety of topics.

In the English 009 essays, written by the same students, the content word *parents* appeared as fifth (2.37% frequency) and *teenagers* as thirteenth (1.37% frequency) in the first set of essays written, while *parents* appeared as third most frequent (2.71%) and *teenagers* as ninth (1.56% frequency) in the second set of essays. This may suggest that the students are using more key words more frequently. However, these were the only two content words used in the top ten most frequent words. Also, in both sets of these essays, the English 101 (mentioned above) and the English 009, when the same students wrote on two different expository topics in two different rhetorical modes or when another set by the same students on a similar topic in the same rhetorical mode written at two different times, the most frequent words were function ones similar to those described in Table 1 in the other English courses, *the* again being the most frequent reaching an average of 5% frequency. In fact, Biber et.al. (1998:29) state that "…function words tend to be the most common words in all corpora" and O'Keeffe and Farr (2003) report in their word frequency lists that *the* ranked in the ten most frequent words in four different sets of corpora examined.

Running through the complete frequency lists of all the English courses, the author noted that very few words were discipline related. Although it may be argued that the essay topics in the present study may not lend themselves to discipline related vocabulary, the range of words was indeed limited even for basic level English considering that the students were of university level. Also, although repetition has been found to have a positive cohesive function in texts (Hoey 1991; Bacha 2002) too many of the same words were being repeated with very few or no synonyms. Furthermore, even though no significance tests at this stage were carried out, four letter words seemed to be the average, showing also limited lexical levels.

As a representative sample of the foregoing, frequency tables are given below (see Tables 2 and 3) for illustrative purposes. The first and last twenty most frequent words are given out of 2232 types of words in the first written essay by the English 009 students (N=36) on the cause-effect topic and the same for the second writing out of 2057 types by number of occurrences and percentage of the total number of words in the text. In Table 3, all the words indicated appeared only once in the corpus of essays and thus no percentages could be calculated. It can be noticed that words appearing only once are not that 'sophisticated' indicating that between the first writing and second there is very little improvement in lexical sophistication. There seems to be a lot of repetition of a few content 'key' words such as *parents* and *teenagers* in both sets of essays confirming some previous research into lexis in the English 009 writing (Bacha 2002). Also, all the words with very low frequency begin with the letters *y* or *w;* probably indicating that there might be a relationship with the beginning letter of words and level of difficulty. It can also be noted that the content words are not specifically discipline related nor 'sophisticated' words for university level. Further, the subordinators *whom* and *which* are rarely used indicating that perhaps coordination was more prevalent than subordination in these essays, a characteristic feature of L1 Arabic non-native speakers of English (e.g. Zughol & Hussain 1985). The foregoing mentioned features were found to be similar in the all the essays in the corpus in this study.

Table 2: Most Frequent Words in English 009 N=36 Sample

N	**009 – Beginning of Semester N=36**	**Frequency**	**%**	**009- End of Semester N=36**	**Frequency**	**%**
1	The	2,357	5.09	The	2,334	5.78
2	And	1,729	3.74	And	1,445	3.58
3	Of	1,201	2.60	Parents	1,095	2.71
4	To	1,169	2.53	To	1,001	2.48
5	Parents	1,097	2.37	Their	8.16	2.02
6	In	883	1.91	A	765	1.89
7	Their	827	1.79	Of	742	1.84
8	A	796	1.72	In	732	1.81
9	Are	747	1.61	Teenagers	630	1.56
10	They	691	1.49	Are	576	1.43
11	Is	658	1.42	They	516	1.28
12	That	651	1.41	For	514	1.27
13	Teenagers	634	1.37	That	499	1.24
14	For	565	1.22	Is	478	1.18
15	Children	557	1.20	Children	455	1.13
16	Between	432	0.93	With	425	1.05
17	Disagree	406	0.88	Between	386	0.96
18	Have	402	0.87	Would	365	0.90
19	As	374	0.81	His	323	0.80
20	This	349	0.75	On	319	0.79

Table 3 Least Frequent Words in English 009 N=36 Sample

N	**009 –Beginning of Semester N=36**	**N**	**009- End of Semester N=36**
2213	Wives	2038	Wealth

2214	Wise	2039	Weapon
2215	Wise-man	2040	Wear
2216	Wiser	2041	Western
2217	Would	2042	Whom
2218	Woman	2043	Which
2219	Women	2044	Willing
2220	Won’t	2045	Wishes
2221	Worked	2046	Witness
2222	Workers	2047	Would
2223	Working	2048	Woman
2224	Works	2049	Women’s
2225	Worldwide	2050	Worked
2226	Worried	2051	Worth
2227	Worth	2052	Wouldn’t
2228	Write	2053	Year
2229	Writers	2054	Yes
2230	Written	2055	Younger
2231	Young	2056	Youngsters
2232	Younger	2057	Youth’s

The reader skimming through the two representative sample essays, selected at random for illustrative purposes (see appendix), can easily discern the limited vocabulary of two different students’ writing whether they are from the basic English course (009) or the more advanced course (202). Even when the same students wrote on the same topic and in the same genre as in the sample N=36 in the English 009 course, the limitation of the lexis is also just as apparent (see Tables 2 and 3). In contrast, the scanned passages from the required textbooks (see appendix) indicate more sophisticated content words which is probably to be expected considering that they are being compared with learner texts.

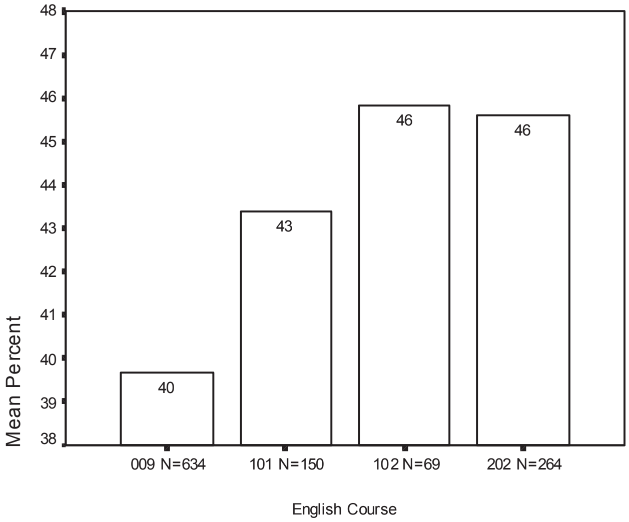
6.2. Word repetition (research question 2)

The results of this research question are given based on the standardized type/token ratio in the essays in six main parts as detailed below.

1. Figure 1indicates that although there is an increase in the type/token ratio from the basic course to the advanced, there is no great difference. It seems that there is quite a great deal of repetition of the same words even in the advanced English 202 course. It is not unusual to find only 40% of different words in remedial English essays; however, 46% does seem to be quite a low percentage for the more advanced essays. When an equal number of essays (N=69) was computed for all English courses, results were comparable. A possible explanation for repetition to be continually high over the courses is perhaps the comparison is being made among courses with different objectives with different expectations, the latter being higher in the more advanced course and thus the

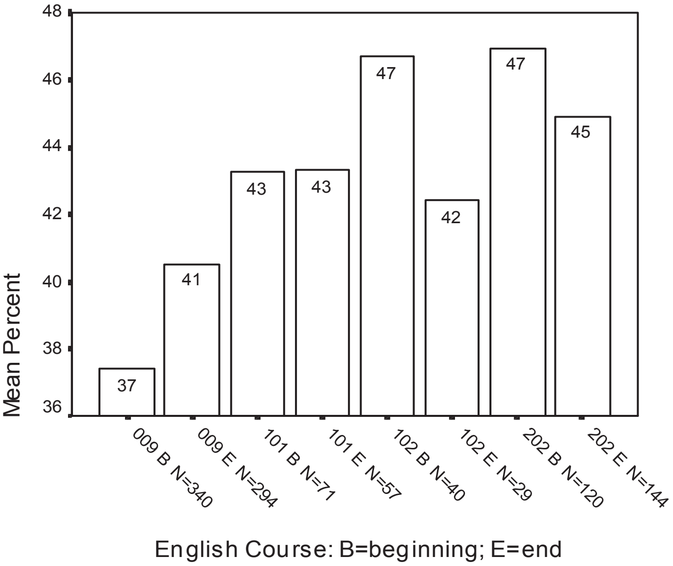
added difficulty for student writers. Also, the English 202 essays were in a different rhetorical mode and this may have affected the results. However, although only two, when one reads through the basic and advanced student sample essays (see appendix) they both exhibit lexical limitation.

Figure 1 Standardized Type/Token Ratio in English Essays as a Whole



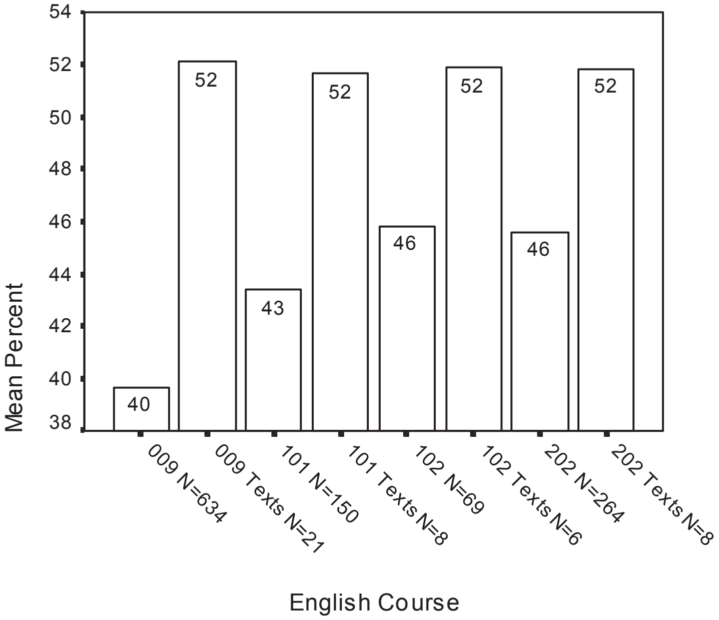
3. Figure 2 shows very little difference between beginning and end of semester essays. The greater difference is indicated in the 009 essays, none in the 101, and a decrease in the end essays in 102 and 202. Again, the results are descriptive and no generalizations can be made. However, if we take the results at face value, there seems to be very little vocabulary development over the course of the semester.

Figure 2 Standardized Type/Token Ratio in English Essays at Beginning and End of Semester



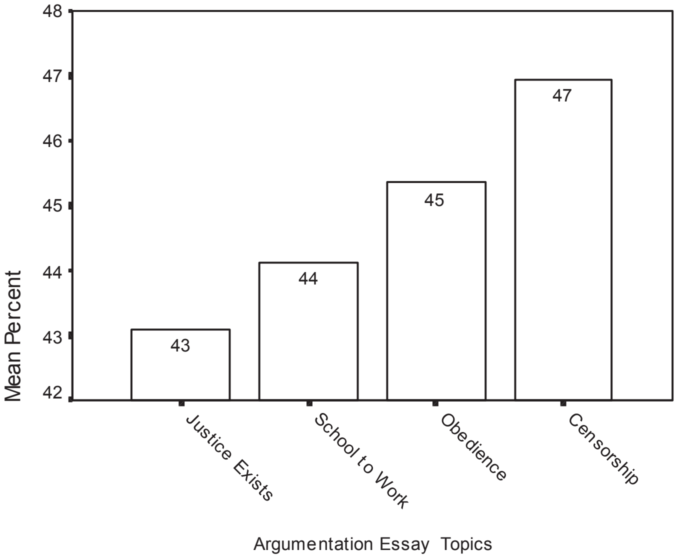
3. Figure 3 indicates that when each of the English course essays were compared to the scanned passages from the respective English course textbooks, there is a difference, the latter showing a higher percentage of different vocabulary in all courses. However, one would expect that the passages from the textbooks would have higher percentages. Perhaps this is due to using the same norming type/token figure of 500 for the scanned texts. The researcher did, therefore, an analysis using every 1,000 words and found that the results were almost comparable. However, on reading the sample scanned texts (see appendix), it is clear that another factor is involved and that is the sophistication of the lexical items which is not prevalent in the student essays.

Figure 3 Standardized Type/Token Ratios in English Essays and Textbooks



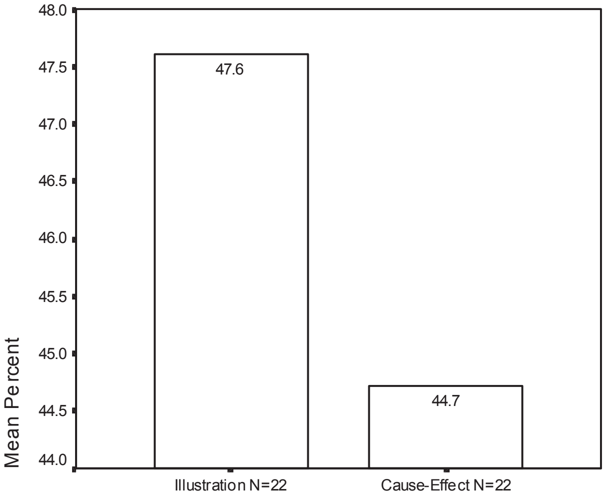
4. Figure 4 indicates that when only the English 202 essays were examined, differences were found in type/token ratio according to the topic. It seems that the topic on *censorship* was perhaps ‘easier’ for the students in the sense that they may have had the vocabulary repertoire to write on such a topic, whereas the *justice* topic seemed quite taxing.

Figure 4 Standardized Type/Token Ratios in English 202 Argumentation Essays



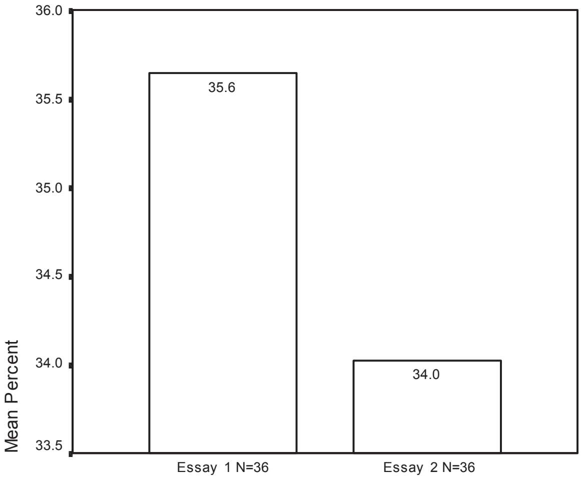
5. Figure 5 indicates that when only the English 101 essays written by the same students on two topics and two rhetorical modes were analyzed, the type/token ratio was comparable with all previous results. It does seem that students, irrespective of genre or topic, do have a problem with a limited lexical repertoire.

Figure 5 Standardized Type/Token Ratio for English 101 Essays by Same Students



6. Figure 6 confirms the above results. Since research has indicated that genre, topic and participants affect results and conclusions made, these variables were accounted for. Figure 6 indicates that there was very little difference in the type/token ratio in essays written by the same students on the same topic and in the same rhetorical mode at two different times. In fact, the ratios on the second set of essays written at the end of the semester decreased in different types of words indicating more repetition of the same words.

Figure 6 Standardized Type/Token Ratio for English 009 Cause-Effect Essays on Same Topic and by Same Students



Although rigorous analysis and testing procedures need to be done, the initial descriptive analysis does seem to confirm the teacher comments that students' lexical repertoire needs to be expanded in writing. More than half of the student essays are repetitive in vocabulary use, and the type of words could be more of an academic sophisticated nature.

7. Limitations of the Study

Although the strength of the present study is that the data was obtained from a natural setting (Biber et.al. 1998), the main limitation is that the essays were those that the teachers were able to pass on to the researcher. Thus, often the numbers in any one rhetorical mode and/or topic were not sufficient to do a large scale sampling in the genre and/or topic and thus the research had to be limited to whole sample analysis and interpretation in most cases according to English course. Thus, random sampling was not possible and thus statistical testing was not done. All the essays obtained were analyzed. Also, the investigation was limited to specific areas and thus only some aspects of the students' lexical repertoire were reported on. More samples of students' essays should be obtained on the same topics and in the same rhetorical mode by the same students since research supports the influence of genre in text analysis. It would also be interesting to follow up on beginning and end of semester essays in a more rigorous experimental research design and also add further dimensions in comparing the obtained student corpus to native-speaker writing as well as other genres written by students in the various disciplines of the university. It is hoped that the corpus now available will be exploited and added to in further studies toward better understanding our students' writing.

8. Implications for L2 Classrooms

The implications from the foregoing study are far reaching. The results, to a certain extent, have confirmed the teachers' concern. If we, as teachers, know how extensive the vocabulary of our students is then perhaps we can better help them in the learning process by teaching and raising their awareness to the importance of vocabulary in writing. Vocabulary (or lexical) corpora analysis, a valid research method, has opened some new frontiers in the teaching/learning of vocabulary with automation which needs to be exploited in our LAU context. This 'limited' study does not profess to have found answers. It does, however, show the necessity to raise both teachers' and students' awareness of the need to improve the teaching/learning of vocabulary in the L2 classroom. It raises questions as what texts to use, what methods to implement and, above all, how students learn vocabulary in the context of their discipline related studies. Answers to these questions cannot be easily found. However, we as teachers could consider alternatives. For example, the texts used in the English courses could be supplemented with discipline related reading texts and/or texts similar to those by Burgmeier, Eldred & Zimmerman (1991) which focus on more specific academic vocabulary instruction. These could be further supplemented with models of discipline related readings to give the students more content discipline oriented vocabulary which was lacking to a great extent in the frequency counts. Last, but very importantly, teachers could also carry out their own individual research in their classrooms which corpora analysis and computer software have made possible (e.g. Coniam 1997).

9. Conclusion

This study was a description of the most frequent words and the rate of repetition in the expository and argumentation texts of students in the LAU EFL Program. Basically, the results indicated that function words are mostly prevalent in these general academic texts and that content words do not begin to become frequent on the list of most frequent words until about the 50th number. Also, repetition, although not a negative aspect when its cohesive function is considered, is quite high with approximate 45% different types of words in the essays across the English courses. There being no great differences from beginning to end of semester according to the two variables studied, frequency and type, nor in any of the English courses suggests that these essays are still learners' texts in the sense that their vocabulary acquisition does not seem to have improved and the vocabulary is restrictive. However, the results must be treated with caution as a more rigorous experimental design needs to be done in order to make any generalisations. Irrespective, the trend seems to be that the lexical repertoire of these students needs to be expanded. To sum up, and hope not to sound too repetitious, words are indeed important and are indeed central to writing. The major problem, therefore as de Beaugrande (2001:10 in Flowerdew 2003) states, "Our major problem is not so much *bad* English or *incorrect* English, as is often lamented, but rather *insufficient* English". The challenge is, however, to what extent can we, through our understanding of students' texts, help or are, as Conrad (2000) qualifies, "willing" to help our students widen their use of words.

Note: This research was partially supported by a grant from the Lebanese American University

References

Aijmer, K. and B. Altenberg (eds.) (1991). *English Corpus Linguistics.* London: Longman Group Limited.

Atari, O. (2004). 'A text-oriented procedure for corrective feedback in an Arab EFL/EAL writing class'. *International Journal of Arabic-English Studies,* 5:207-220.

Bacha, N. N. (2000a). 'Academic writing in a multilingual context: a study of learner difficulties'. *International Journal of Arabic-English Studies*, 2/2: 239-268.

_______(2000b). 'Faculty and EFL students' perceptions of students' language abilities at the Lebanese American University, Byblos Branch'. Unpublished survey results, Byblos: Lebanon.

_______(2001). 'Student perceptions on the relation between vocabulary and academic essay writing at the Lebanese American University.' Unpublished survey results, Byblos, Lebanon.

Bacha, N., M. Cortazzi, and F. Nakhle. (2002). 'Academic lexical literacy: investigating the cohesion of Arabic speakers' essays in English'. *International Journal of Arab-English Studies*, 3/1,2:119-152.

Behrens, L & Rosen, L. (2004). *Writing and Reading across the Curriculum.* New York: Addison Wesley Publishers.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge:

Cambridge University Press.

_______(1993). ‘Representativeness in corpus design’. *Literary and Linguistic Computing*, 8:1-15.

Biber, D., S. Conrad and R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, D. and S. Conrad (2001). ‘Corpus-based research in TESOL: quantitative corpus-based research: much more than bean counting’. *TESOL Quarterly,* 35/2: 331-336.

Boswood, T. (ed.) (1997). *New Ways of Using Computers in Language Teaching,* Alexandria, VA: TESOL.

Boguraev, B. and J. Pustejovsky (eds.) (1996). *Corpus Processing for Lexical Acquisition.* Mass.: The MIT Press.

Bowker, L. (1999). ‘Exploring the potential of corpora for raising language awareness in student translators’. *Language Awareness*, 8:3/4:160-173.

Buscemi, S. and C. Smith (eds.) (2002). *75 Readings Plus*. New York: N.Y. McGraw-Hill.

Burgmeier, A., G. Eldred, and C. B. Zimmerman (1991). Lexis*, Academic Vocabulary Study.* New Jersey: Prentice Hall Regents.

Coniam, D. (1997). ‘A practical introduction to corpora in a teacher training language awareness program’. *Language Awareness,* 6/4:199-207.

Connor, U. (1987). ‘Research frontiers in writing analysis.’ *TESOL Quarterly*, 21/4677-695.

Conrad, S. (1999). ‘The importance of corpus-based research for language teachers’. *System*, 27:1-18.

________ (2000). ‘Will corpus linguistics revolutionize grammar teaching in the 21st century’? *TESOL Quarterly,* 34/3:548-560.

Crusius, T. W. (1989). *Discourse: A critique and synthesis of major theories*. New York: The Modern Language Association of America.

Egbert, J. and E. Hanson-Smith. (eds.) (1999). CALL *Environments: Research, Practice, and Critical Issues*. Alexandria, VA: TESOL.

Flowerdew, L. (2003). ‘A combined corpus and system-functional analysis of the problem-solution pattern in a student and professional corpus of technical Writing’. *TESOL Quarterly,* 37/3:489-511.

Francis, G. (1994). ‘Labelling discourse: an aspect of nominal-group lexical Cohesion’. In Malcolm Coulthard (ed.), *Advances in Written Text Analysis, 83-101.* London: Routledge.

Garside, R., G. Leech and A. McEnery. (eds.) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman Limited.

Gledhill, C. (2000). ‘The discourse function of collocation in research article introductions’. *ESP* Journal**,** 19:115-135.

Grabe, W. and R. Kaplan (1996). *Theory and Practice of Writing.* London: Addison Wesley Longman Limited.

Green, C. et.al. (2000). ‘The incidence and effects on coherence of marked themes in interlanguage texts: a corpus-based inquiry'. *ESP Journal*, 19:99-113.

Halliday, M. A. K. et.al. (2004). *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Huckin, T, M. Haynes, and J. Coady. (1995). Se*cond Language Reading and Vocabulary Learning.* New Jersey: Ablex Publishing Corporation.

Johns, A. (1981). 'Necessary English: a faculty survey'. *TESOL Quarterly,* 15/1:51-57.

Jordan, R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers.* Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Longman.

Hunston, S. and G. Francis (2000*). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

________. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Kroll, B. (ed.) (1990). *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University Press.

Langan, J. (1997). *English Skills with Readings.* New York: McGraw Hill Publishers.

Lewis, M. (2002). *The Lexical Approach: The State of ELT and a Way Forward.* Australia: Thomson/Heinle.

Malvern, D. et. al. (2004). *Lexical Diversity and Language Development: Quantification and Assessment.* Houndmills, Basing Stoke, Hampshire: Palgrave Macmillan.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers.* Cambridge: Cambridge University Press.

McCarthy, M. and R. Carter (2001). 'Size isn't everything: spoken English, corpus, and the classroom'. *TESOL Quarterly*, 35/2:337-340.

McEnery, T. and A. Wilson (1996). *Corpus Linguistics.* Edinburgh, Scotland: Edinburgh University Press.

McWhorter, K. (2003). *Successful College Writing*. New York: St. Martin's Press.

Mukattash, L. (2003). 'Towards a Methodology for Teaching English to Arab Learners (TEAL)'. *International Journal of Arabic-English Studies,* 4:211-234.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

O'Keeffe, A. and F. Farr (2003). 'Using language corpora in initial teacher education: pedagogic issues and practical applications'. *TESOL Quarterly*, 37/3:389-418.

Ravelli, L. J. and R. A. Ellis (2004**).** *Analyzing Academic Writing: Contextualized Frameworks*. London: Continuum.

Rogers, H. (2005). *Writing Systems: A Linguistic Approach.* Malden, Mass.: Blackwell Publishers.

Sa'Addedin, A.M. and M. Akram. (1989). 'Text development and Arabic-English negative interference'. *Applied Linguistics*, 10/1:36-51.

Scott, M. (1998).*Wordsmith Tool Manual*. Version 3.0. [Computer software]. Oxford: Oxford University Press.

Seliger, H., and E. Shohamy (1989). *Second Language Research Methods.* Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

_____ (1994). 'Trust the text.' In Malcolm Coulthard (ed.), *Advances in Written Text Analysis,*12-25. London: Routledge.

Sinclar, J. (2004). *Trust the Text: Language, Corpus and Discourse.* New York, N.Y.: Taylor and Francis.

Simpson, R. and D. Mendis (2003). 'A corpus-based study of idioms in academic speech'. *TESOL Quarterly,* 37/3:419-441.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture.* Cambridge: Massachusetts, Blackwell Publishers.

Zughoul, M. and R. Hussain (1985). 'English for higher education in the Arab world - a case study of needs analysis at Yarmouk University'. *The ESP Journal*, 4:133-152.

Appendix – Sample Essays

009 End of Semester Student Essay - Contrast

The buying habits of high income families are different from the ones of low income families.

The high income families by a new model as soon as the old one fails , for instance if the toaster is broken , they don't fix it, even when they know that it can be fixed , because they have the money to buy a new one , so why they bother ?

In addition to that they have expensive things , they can posses a new model car , expensive clothes , and expensive furniture's, as well , they don't care about the price , but they are victim of the competitive urge , they need to have this and that ; if they want to install a swimming –pool or change the house , they can because they have the money to compete , this is what makes them victims of competitive urge .

However , the low income families recycle the things that can be reused and fix the broken ones like the heater , the bicycles, and so on . In addition to that low income families cant change furniture every year or change the car or buy a boat , so they cant be victim of the competitive urge but even if they were they can't compete because they don't have enough money. So buying habits of high income families are different from the habits of low income families. (244 words)

202 End of Semester Student Essay - Argumentation on Justice Exists

Since the antic Greece, when Plato or Cicero where sacking for a fair and judicial is one of the most important key to settle a democratic and constitutional political system although, at the beginning of the new millennium some citizens, in democratic country as France as united sates believe that justice exists protecting their rights, many people clams that there is no justice due to loopholes to out coming from the justice system itself.

Who can say that justice doesn't exist? Justice is ruled by fundamental laws, a men demands and legal constitutions, which are the pillars of democracy. Those who believe in democracy consequently should admit there exist justice. If the justice not exists, then would the jails be full as they actually are? "None is above the law doesn't mean there is a justice. Are the laws protecting human rights are applied, then it should be a justice in the same way as citizens conceive the existence of democracy even with the smallest accident or a juries, one can invoke justice. Women order 49 cent coffees from McDonalds, 29 million dollars after it spills on her. But one question remains: what does justice mean? Can one call justice if one innocent person is condemned guilty or vice versa? Mistakes occurs: in rare human est. “but” preserve est. diabolic in other words to continue his mistakes is worth. At it is shown in the "when the police blunder a little", by Bernett Beach their exist some loopholes in the juridical system. Will the same laws, which can condemn people, a person, who is guilty, can be freed.

Another illustration of this injustice is pictured in the movie primal fear where a young murder is found innocent due to his acting thermal performance. He makes believe that he is psychologically sick. More recently, a famous dictator in Chily, the general Cinoche, succeeded to avoid any punishment due to several civil massacres, by faking and abusing the juridical system.

It may be true those crimes, and out laws are condemned and punished by a justice, however the system is not perfect. It allows criminal to escape from a complete fair judgment. The juridical system must be improved and modified to tend towards a more. Perfect justice even if one knows that justice is blind. (384 words)

009 Part of a Scanned passage from Textbook - Reasons

My Car Accident (Langan 1997)

1Several factors caused my recent car accident. 2First of all, because a heavy snow and freezing rain had fallen the day before, the road that I was driving on was hazardous. 3The road had been plowed but was dangerously icy in spots where dense clusters of trees kept the early morning sun from hitting the road. 4Second, despite the slick patches, I was stupidly going along at about fifty miles an hour instead of driving more cautiously. 51 have a dare devil streak in my nature and sometimes feel I want to become a stock-car racer after I finish school, rather than an accountant as my parents want me to be. 6 A third factor contributing to my accident was a dirty green Chevy van that suddenly pulled onto the road from a small intersecting street about fifty yards ahead of me. 7The road was a sheet of ice at that point, but I was forced to apply my brake and also swing my car into the next lane. 8 Unfortunately, the fourth and final cause of my accident now presented itself. 9 The rear of my Honda Civic was heavy because I had a barbell set in the backseat. (201 words)

202 Part of a Scanned Passage from Textbook - Argumentation
Why We Crave Horror Movies (Behrens & Rosen 2004)

To think of modern horror fiction is to summon the name of Stephen King. Author of such best-selling novels as Carrie: A Novel of a Girl with a Frightening Power (1974), The Shining (1977), Pet Sematary (1983), and Misery (1987), King has devoted a career to exploring our nightmares and making them come alive. His novels sell in the millions; the movie adaptations based on them play to packed (screaming) houses-all testament to King's mastery of a form that prompts a simple but mystifying question: Why do people pay good money to be scared? Over his career, in various interviews and essays, King has observed that we seek out and respond to horror in fiction as a strategy for contending with the horrors and insanity of our daily lives. In the essay that follows, he observes how a good horror story lets us keep the "alligators" lurking in our psyches fed. The premise is clear: each of us maintains both a civilized, public face and then something altogether nastier that we keep hidden but must nonetheless "feed. " Good horror stories and movies do just that. (186 words)