MELTA

Malaysian Journal Of ELT Research

ISSN: 1511-8002

Vol.3, 2007

Investigating Spelling Errors In A Malaysian Learner Corpus

Simon Botley @ Faizal Hakim
Doreen Dillah

Universiti Teknologi MARA Sarawak, Malaysia

Abstract

This paper describes an empirical study of spelling errors using a learner corpus of university-level English. The corpus, known as CALES (Corpus Archive of Learner English in Sabah/Sarawak), consists of argumentative essays collected from university students in three public universities in Sarawak and Sabah. After describing the methodology of the CALES project, the paper outlines how spelling errors can be classified, using a combination of pre-existing categories from the literature, and categories observed in the data. The data demonstrates clearly that spelling is still a major issue both for teachers and learners, and that many students make spelling errors that fit into known categories, despite the fact that they have been studying English for at least 10 years in the Malaysian education system. As well as making observations, this paper offers a number of speculations about why students make spelling errors, and proposes some recommendations of how to prevent these errors from appearing in student writing.

Introduction

Spelling errors are highly ubiquitous and contentious features of Second Language learners' written performance. They are ubiquitous because despite years of drilling and training in school and university, spelling errors still appear in large numbers in the writing produced by learners. Spelling errors are contentious, therefore interesting, because they reveal a great deal of information about three key aspects of the language learning process. Firstly, spelling errors can tell us much about a student's interlanguage (Selinker, 1972). Interlanguage refers to the type of language produced by learners, which reflects the state of proficiency attained in the target language at a given time.

To paraphrase James (1998:5) the interlanguage (IL) can be seen as the learners' version of the Target Language (TL), and in effect describes point the learners have reached on their route from total ignorance of the TL to proficiency and mastery in it. Once we know some details of the learners' IL, this information can be used by educators to target those aspects of the target language which require more input, and can tell researchers much about the process of language learning.

Secondly, spelling errors reveal and reflect the influence of the first language on the learner's target language performance. As is shown in the data analysed in this paper, the learner's first language spelling rules can and often do get mis-applied by learners when it comes to turning phonemes into graphemes in the target language. And finally, spelling errors may even provide clues as to the cognitive state of learners, especially where there is a need for clinical intervention. Examples may be dyslexia, where cognitive damage or illness may cause characteristic spelling errors to occur, which require therapy to eliminate them.

This paper will focus on spelling errors from the perspective of Error Analysis (EA). Error Analysis was defined by Richards and Schmidt (2002:184) as the study and analysis of errors made by second language learners, with a view to identifying language learning strategies, identifying the causes of errors and identifying areas of difficulty for language learners. James (1998:5) points out that EA is a paradigm that involves objective analysis of the interlanguage of learners and a comparison of the IL with the TL, to see where the two differ. As we will see below, this often involves some consideration of the learners' mother tongue, although EA was originally set up as a distinct enterprise from Contrastive Analysis (CA), where the mother tongue was compared with the IL, in order to predict likely errors.

In this paper, more specifically, we are working within a newer paradigm, or a newer flavour of EA, known as Computer-aided Error Analysis – CEA (Granger et al, 2002: 11-14). CEA involves analysing errors in a computer learner corpus, whereas traditional EA did not make use of computer corpus data. CEA uses computer technology in the form of specialized search and retrieval software such as Wordsmith Tools (Scott, 1996). Also, it allows more standardisation in the analysis of errors as well as discussing errors in

context. Importantly, CEA is capable of analysing very large amounts of data stored digitally in a computer corpus.

CEA can be carried out in two ways, according to Granger et al (2002:11-14). Firstly, the analyst can scan through a learner corpus for all examples of a chosen error-prone linguistic feature, using text-retrieval software such as a concordancer. This is relatively fast, but might be limited only to those features that the analyst thinks are problematic. Furthermore, in carrying out CEA in this way, an analyst may miss out important features which are not initially thought of as problematic.

The second method of carrying out CEA is to apply a pre-designed set of error categories, or tags, to a learner corpus, and identifying and classifying all examples of errors in a corpus. Although this method is highly labour-intensive (see Dagneaux et al, 1998 as well as Botley et al, 2005), it is a very powerful method of revealing large numbers of problematic features of the interlanguage that may not have been predicted by the analyst, as well as many that have been predicted. This type of analysis can be made easier by using specially-designed software such as the UCLE Error Editor used in this paper (Dagneaux et al, 1998).

In this paper, then, we adopt a CEA methodology to investigate spelling errors identified using error tags. The data for this study comes from a new learner corpus, called CALES (Corpus Archive of Learner English in Sabah/Sarawak), which is under development in UiTM Sarawak and Universiti Malaysia Sarawak (UNIMAS), with some input from Universiti Malaysia Sabah (UMS) and UiTM Sabah.

Before describing the study proper, it is necessary to firstly outline how the corpus came about and how it is structured. After this, we will explore some basic concepts in the study of spelling errors within an Error Analysis (EA) framework. Finally, we will present some findings from CALES concerning the different categories of spelling error identified in the data, along with some frequency statistics and examples.

The CALES Learner Corpus

What Is A Learner Corpus?

A learner corpus is a computerised text database containing spoken or (primarily) written material produced by second-language (L2) learners. Although any collection of student written material gathered together by teachers can be considered a learner corpus, such a collection is not considered a corpus proper unless it is planned and collected according to clear and sound design principles.

Such design principles help the corpus to be a representative sample of the work produced by the students under analysis. It would be necessary to collect the material so

that it reflects various characteristics of the learners themselves, and the learning situation.

For instance, Granger, Dagneaux and Meunier (2002:13), in Barlow (2005:338-339), divide variables which may be relevant in designing a learner corpus into Learner Variables and Task Variables. Learner Variables may include age, learning context, proficiency level, etc., while Task Variables can include medium (written or spoken), field (e.g. general, current affairs etc.), genre (argumentative or narrative) and length (in words).

It would be possible to add more variables, as is done in the CALES project, and these variables can help the corpus builder to design an appropriate instrument to capture information relevant to them. A learner corpus, however it is designed, can be used for many purposes, but a common application is in the investigation of the features of language used by students who are learning a new language (the students' interlanguage (IL) (Selinker, 1972). This means that a learner corpus offers teachers and researchers a great deal of valuable information about the errors that learners make when they attempt to use their target language during the language learning process.

Learner corpus linguistics has developed into a well-defined field of research in recent years, and the prominent learner corpus project has been the International Corpus of Learner English (Granger, 1998 and Granger et al, 2002) which is based at the Université Catholique du Louvain in Belgium.

The ICLE is a large publicly available learner corpus containing essays written by university undergraduates studying English in primarily European universities. The ICLE consists of essays of approximately 700 words in length with an aim to produce national sub-corpora of approximately 200,000 words per country (Barlow, 2005:338). There are ICLE component corpora from over 17 countries including Germany, Poland, Sweden and Russia, with plans to extend the coverage beyond Europe to China, Japan and Brazil.

Alongside the ICLE, learner corpus work has taken many forms and has had many different aims. These include designing and analyzing corpora and software tools (Meunier, 1998), corpus-based studies of grammar, lexis and discourse (Ringböm, 1998), 'interlanguage' studies (Altenberg, 2002), and finally dictionary design and textbook development (Kaszubski, 1998).

Furthermore, as is pointed out by Granger et al (2002:22-26), the findings derived from learner corpus research can even be used in curriculum design (Granger et al, 2002: 22-24), materials design, (Milton, 1998) and classroom methodology. (Granger and Tribble, 1998). Here in Malaysia, learner corpus linguistics has also started to make its presence felt, with three main projects already completed or under way. These are the EMAS (English of Malaysian School Students) corpus (Arshad et al, 2002), the ongoing MACLE (Malaysian Corpus of Learner English) (Knowles and Zuraida, 2004) and

CALES (Corpus Archive of Learner English in Sabah-Sarawak) (Botley et al, 2005, 2007).

The EMAS corpus is a very large and demographically sampled corpus of student writing and speech collected in primary and secondary schools all over Peninsular Malaysia. The MACLE corpus is still in development and aims to be a future Malaysian sub-component for the ICLE. Finally, CALES, which will be described in the next section, is aimed at complementing the MACLE, and is collected in Sarawak and Sabah. The data used in this paper is derived from the CALES corpus.

The CALES Corpus

The CALES corpus is an ongoing project which was started in 2003. At the time of writing, about 400,000 words of argumentative essays have been collected from students taking English proficiency courses at UiTM's Sarawak and Sabah Campuses, Universiti Malaysia Sarawak (UNIMAS) and Universiti Malaysia Sabah (UMS).

The CALES corpus followed as closely as possible the methodological and design principles of the International Corpus of Learner English (ICLE) (Granger, 1998; Granger et al 2002). Students wrote argumentative essays in class under timed conditions on a range of topics. The essays were supposed to be up to 1000 words in length but in the end ranged from 200-800 words.

As well as writing essays, the learners were asked to complete a Learner Profile instrument which gave details about the learner (age, race, language background, English proficiency etc) and the task (topic, length, setting etc). Essays were digitized by skilled typists and the data from the Learner Profiles were entered into a database. A sample essay is shown in Appendix 1, and a list of the essay topics is included in Appendix 2. The essays were stored in separate digital files which could be sampled and analysed to reveal findings about spelling errors. However, before discussing the methodology and findings for this study, let us explore some basic concepts.

Spelling Errors And Error Analysis

An excellent account of spelling errors within the context of Error Analysis (EA) is provided by James (1998: 129-139), who distinguished between ‘mis-spellings’ on the one hand, and ‘mechanical errors in writing’ on the other. Both mis-spellings and mechanical errors are for James classified as ‘substance errors’, meaning that they are related to the medium utilised by language users – either written or spoken. According to James, mis-spellings and mechanical errors are caused when a learner makes an encoding error while writing.

James identifies four types of 'mechanical errors', namely punctuation errors, typographical errors, confusables and dyslexic errors. Punctuation errors involve all commonly known errors in using punctuation marks and spacing in written texts. They include under-use and overuse of punctuation marks, for instance 'a boys club' or 'tomato's', splits (Carney, 1994:84) such as 'to gether' or 'an other', and fusion, for instance 'takeaway' or 'cashpoint'.

Typographical errors are primarily caused by mis-keyings committed by typists, and differ from other spelling errors in that they appear only in printed or typed text. Such 'typos' are caused not by linguistic ignorance or memory slips, but rather by simple mechanical clumsiness in operating a complex machine at speed. James (1998: 131) points out that many of these errors are caused by a typist striking a key that is adjacent to the correct one on a QWERTY keyboard, e.g. 'tge' instead of 'the', and also include such common proofreader's banes as reversals ('adn' for 'and'), omissions ('lenth' for 'length') and anticipations ('extexted' instead of 'extended').

Confusibles are confusions between word pairs that have similar-sounding phonemes or morphemes, such as 'divorce/devoice', 'anus/onus', 'course/coarse' or 'discrete/discreet'. Confusibles were described by Carney (1994:82) as 'phonetic near-misses'. These often give rise to what are known as 'malapropisms', named after Sheridan's literary character Mrs. Malaprop.

The final class of mechanical error identified by James is the dyslexic error, which refers to errors made by language users who may suffer from a pathological condition, such as aphasia or dyslexia, which impairs their language production. James (ibid: 133) points to three main dyslexic spelling phenomena, and provides a number of examples. Firstly, there is mis-selection of two letters that can represent the same sound, as in 'parc' versus 'park'. Secondly, there are mis-orderings of letters, as in 'tow' versus 'two'. And finally there are letter-reversals, or 'strephosymbolia', commonly observed in the writing of dyslexics, as in 'adowt' versus 'about'.

Returning to mis-spellings proper, James (ibid: 134) defines mis-spellings thus: "Misspellings (MSs) as such violate certain conventions for representing phonemes by means of graphemes". While phonemes are the minimal units of sound that carry meaning in linguistics (for instance /p/, /b/ in 'pit' and 'bit' – in this paper, we use the common linguistics convention of enclosing phonemes with / / marks), graphemes are the smallest written feature which can carry meaning (<p> in 'pit' and <b> in 'bit – here, we follow James' use of the < > symbols to mark graphemes). Therefore for James, mis-spellings occur when the rules, often known as phonographic (PG) rules, that determine how a given phoneme is to be represented in writing, are broken.

In an earlier study, James et al. (1993) investigated spelling errors in the writing of primary school children whose dominant language was Welsh. Although this study made a number of different observations about spelling, the most pertinent for our purposes are

that James and his co-workers derived two broad categories of mis-spellings from the Welsh children's writing.

The first category was 'mispronunciations'. These are caused by a basic mispronunciation of a target sound in the second or target language (L2), resulting in a learner choosing the wrong grapheme to represent that sound. James (1998:137) uses a Welsh example to illustrate this. In trying to spell the English word <blood>, Welsh school children substituted the target phoneme [ʌ] for the Welsh equivalent [ə] (here, phonetic sounds are represented using square brackets) which in Welsh is represented by the letter <y>. The result of this was that the word 'blood' was frequently mis-spelled <blyd>.

Another example closer to home is the Malaysian spelling of the English word 'phenomena' as <fenomena>, which occurs in the CALES data. Here, the student writer has most likely mis-applied a Malay language PG rule ("only represent the sound [f] with the letters <f>, never <ph>"), and ended up with <fenomena>.

The second category of mis-spellings identified by James et al (1993) is 'written misencodings', which are not caused by pronunciation. Written misencodings are divided into interlingual and intralingual types. Firstly, interlingual misencodings lead to spelling errors that can be linked back to the learner's L1 (first language). James et al (1993) further subdivided these into three types, which will be briefly described here, with Malay examples where possible. The first subtype occurs when a learner tries to use a spelling rule from their L1 which does not exist in the target language. Two examples from Malay illustrate this sub-type quite well. Firstly, the English words 'cent' and 'graph' may be rendered <sen> and <graf> as a result of L1 influence. Furthermore, it should be noted that these examples are borrowed lexical items, and the Malay language will apply its own PG rules to borrowed items like these, with the result that when students write these words in English, they may retain the Malay spellings, hence the misencoding.

The second interlingual subtype is where a learner attempts to use a grapheme from his or her L1 which also exists in the target language, but has a different phonetic value in that language. An example of this is the grapheme <c> which exists in English and in Malay but in Malay, this grapheme represents the sound [tʃ] as in 'capak' (neglect) and 'cantik' (beautiful), whereas in English it is used for the [k] and [s] sounds as in 'cap' and 'ceiling'.

The final subtype is more subtle. Here, a learner may use a grapheme that exists in both the L1 and the target language, but that same grapheme is distributed differently in the target language than it is in the L1. James (ibid:138) illustrates this with another Welsh example. In Welsh and English, the sound [f] can be spelled <ph>, as in <ei phen> (Welsh for 'her head') and <phone>. However, while English allows the <ph> to appear word-initially, word-medially and word-finally ('phone', 'nephew', 'graph'), in Welsh, <ph> can only occur at the beginning of words. This, according to James (1998:138),

results in Welsh learners of English making spelling errors such as <neffew> and <graff> because they can only apply their own L1 distribution rules for this particular grapheme.

Finally, intralingual misencodings include a number of different subtypes, according to James et al (1993). The first one is the overgeneralisation of an L1 spelling rule. James (ibid: 138) uses the example that [jə] is written as <iour> in words like 'saviour' but this cannot be generalised to words like 'picture' (resulting in a misspelling like <pictiour>).

The second subtype is 'homophone confusion', seen in confusable pairs such as 'turn/tern', 'roll/role' and 'their/there'. Thirdly, there is 'mis-choice' where a learner simply chooses the wrong grapheme, resulting in a word that does not exist in the target language, although it could, e.g. spelling the word 'mean' as <meen>. Finally there is 'letter naming' which is a spelling strategy involving using letters to represent sounds that are the same as the sound of the name of the letter in question. A good example of this is SMS language where users of the service may write <mt> to represent 'empty' or <c u> to represent 'see you'.

James' work has given us a number of useful categories which can help us to investigate spelling errors in a corpus. James' categories can be divided firstly into Mechanical Errors, which include Punctuation, Typographic, Confusibles and Dyslexic errors. Secondly, there are the Mis-spellings, which are divided into Mispronunciation Errors and Written Misencodings. Written Misencodings in turn can be further divided into Interlingual and Intralingual. Interlingual written misencodings can use L1 spelling rules that do not exist in L2, can use L1 graphemes with different L2 phonetic values, or can use L1/L2 graphemes with different L2 distributions. Finally, the Intralingual written misencodings include overgeneralization, homophone confusion, mis-choice and letter-naming.

Now that we have some idea of the ways in which spelling errors can be classified, we will examine some data extracted from the CALES learner corpus.

Methodology

Data Analysis

We will now describe a study which was carried out on the CALES data to investigate spelling errors. To this effect, two samples from the CALES data were selected, consisting of 135 essays written by Degree students and 146 essays by Diploma students. The details of the distribution of these files by state origin and institution are given in Tables 1 and 2 below.

The information which enabled the essays to be sampled like this was taken from the Learner Profile database constructed from the LP returns. Furthermore, the Diploma-

level data was collected from UiTM Sarawak only, hence the lack of different columns for institutions in Table 2.

Table 1: Degree-Level Files In The Spelling Error Study

	UiTM	UNIMAS	UMS	*TOTAL*
Sarawak	48	0	3	51
Sabah	32	5	12	49
Semenanjung	11	0	18	29
Others	0	0	6	6
TOTAL	91	5	39	135

Table 2: Diploma-Level Files In The Spelling Error Study (Uitm Sarawak Only)

Sarawak	Sabah	Semenanjung	TOTAL
100	2	44	146

As can be seen above, students from Sabah and Sarawak provided more essays than Semenanjung students, and there was a preponderance of data from UiTM. This reflected the state of the corpus collection at the time of this study, and it is hoped that the CALES corpus will become more balanced in terms of institution and state of origin as more data is added. All of the files selected for the study were firstly annotated using the UCL Error-Tagging scheme (Dagneaux et al, 1998). As only spelling errors were the focus of this study, only one error tag was required - the tag 'FS' (Formal, Spelling). This tag was manually applied using a text editor to all examples of spelling errors identified in the sample essays.

Following the UCL tagging methodology, 'correct' target spellings were entered (surrounded by $ signs) next to the incorrect words, for instance 'fenomena FS $phenomena$'. Once this was done for all cases of spelling error identified, all files were concordanced using Wordsmith Tools, to gain frequency counts and contextual information to allow spelling errors to be classified and studied further

Findings

In the samples under analysis, 1,018 spelling errors in the Degree-level sample, and 867 errors in the Diploma-level sample, were identified. These errors fell into a number of categories, many of which broadly fit into the framework given above by James (1998). These categories are doubling, omission, addition, mis-ordering, mis-use of punctuation, replacement, L1-influence, US spellings, mis-pronunciations, word coinage and direct borrowing. These categories are discussed below, along with some examples from the corpus. In all examples, the spelling errors are rendered in bold type and enclosed in < > brackets. The examples are typed exactly as they were originally written.

Firstly, the corpus data contained a great deal of **mechanical errors**, especially punctuation and typographical errors. These can be classified into the following, as can be seen in examples (1) to (6):

Doubling:

(1) 'As a result the richers people like <**abbuse**> the poor people.'

Omission:

(2) '…microwave to cook dinner washing machine to wash cloth and <**vacum**> to clean the house.'

Addition of vowels or consonants:

(3) '..I am defenitely agree that this <**entertaintment**> programmes is a harmful…'

Mis-ordering (including reversals):

(4) '…also some people that just simply give their family eat with this <**frobidden**> money.'

Mis-use of punctuation, spaces, abbreviation, capitalisation etc:

(5) 'It means <**alot**> to us, nowadays.'

And:

(6) 'Apprentice, Fear Factor are some of the good examples of reality <**tv**> programmes which should be watching by Malaysia'

The next category of error identified falls into James' class of mis-spellings proper, although many are difficult to classify rigorously using James' typology. We have called this category **consonant/vowel replacement**, because what happens in most cases is that a particular phoneme gets replaced by something else, which may or may not be motivated by pronunciation issues. Here are three examples from the corpus data:

(7) 'so example where rich people become poorer because the tried to <**abtain**> the money in a wrong way.'

(8) 'the twin tower, the Formula - 1 circuit, one of the <**sofisticated**> circuit in the world'

(9) 'For example robbery, <**prostitude**>, <**drag deller**>'

In example (7), there is some confusion between pairs of vowel sounds [a] and [ɒ], motivated perhaps by the fact that both sounds can be realised in speech by the neutral schwa sound [ə]. In example (8), the learner has simply used the wrong grapheme to realise the fricative sound [f], given that the <ph> grapheme does not exist in Malay. This is what James might term 'homophone confusion'.

Finally, example (9) is very interesting, in that the first word <prostitude> uses a grapheme <d> realising the minimal pair of the correct version <t>. Also, the <drag deller> would appear to be an example of a mispronunciation, perhaps motivated by L1 influence.

The next category of errors in the CALES corpus samples further highlights the influence of the L1. Here, we identified a large number of cases where a learner's spelling of a target language word is heavily influenced by Malay spelling rules, in some cases signaling a direct lexical borrowing which is mis-encoded. Here are some examples:

(10) '..to use computers to do their filing, proposal, contract, designing, <**accaunting**>, key in personal detail..' (L1-influenced spelling)

(11) 'They forget about love, relationship, and <**karier**>' (not a Malay word, but influenced by Malay spelling rules).

Next, there are a few examples of **American spelling** in the data, no doubt reflecting the influence of American spelling conventions on Malaysian learners of English, despite the official use of British spelling in Malaysia and the use of British English in school and university. Here is an example:

(12) 'Therefore, we should control our <**behavior**> and habit…'

Next, there are a small number of cases of errors which resemble mispronunciations, but do not involve replacement of a vowel or consonant. An example would be:

(13) 'We should be proud <**becoz**> our country can have a good performer..'

In example (13), there seems to be a wholesale rendering of the learner's own pronunciation of the word 'because' which may be caused by a lack of awareness of the English spelling, or by a phonetically-motivated misencoding.

Finally, there are many cases of what we term **word coinages**, which would appear to correspond to James' category of **mis-choice**, where the result is a word that does not exist in English, although it is plausible. Here are some examples:

(14) 'Corruption occured everywhere and <**anywhen**>.'

(15) 'It also helps Malaysia's economies develop <**fastly**>.'

(16) '..people dare to do <**unuseful**> activities such as <**robbering**> the bank, <**stoling**> money'

Not all of these examples are exactly cases of mis-choice in James' sense, however. They appear to be motivated by a mis-application of morphological rules, e.g. 'robbering' and 'unuseful'. Also, the form 'stoling' above appears to have a phonological motivation for the **mis-choice**.

Discussion of Findings

We have reported some examples of spelling errors in the CALES corpus, to show how they fall into various categories, some of which are supported by existing literature. To sum up our findings, Table 3 below provides some statistics on the frequency of these categories in the two samples analysed in this study. The table has been sorted to display the most frequent categories overall, in descending order.

Despite the fact that the data sample used in this paper was relatively limited, we can make a number of observations about spelling errors from this data. Table 3 tells us firstly that James' category of mechanical errors are by far the most frequent in this data (about 82% of the total), with a preponderance of omission (accounting for almost 1/3 of the total errors), replacement, addition and mis-use of punctuation in both samples.

In particular, it appears surprising to see a high frequency of such mechanical errors which appear not to be motivated by linguistic considerations such as the L1. An example of this would be simple punctuation errors, many of which seem to involve violation of conventions concerning capitalisation (<dvd>, <sms> etc) and spacing (<eventhough>, <everytime>).

Table 3: Spelling Error Frequencies In Degree And Diploma Level Essays

Category	Degree	Diploma	TOTALS	%
Omission	314	274	588	31
Replacement	184	138	322	17
Addition	140	79	219	11.61
Mis-use of punctuation	106	83	189	10
L1-influence	87	84	171	9
Mis-ordering	65	67	132	7
Doubling	43	51	94	5
Word coinage	32	61	93	4.93
US spellings	30	8	38	2
Direct borrowing	15	7	22	1.17
Mis-pronunciations	2	15	17	1
TOTAL	1,018	867	1, 885	100*

**After rounding up.*

Admittedly, this begs the methodological question of whether these errors are likely to have been the result of mis-keying by the typists who digitised the hand-written essays after they were produced by the students. It is one of the oft-cited shortcomings of CEA that because much of the analysis of the data is manual, this may introduce a certain degree of subjectivity in the analysis, despite the presence of automatic techniques to count and sort various errors.

To address this concern, we may turn again to Dagneaux et al (1998), who highlighted the value and desirability of such manual analysis in CEA. In their own detailed manual error analysis study, Dagneaux et al (ibid) demonstrated that manual methods are inevitable, given the current lack of automatic methods of capturing errors in handwritten scripts. Technologies such as character recognition and optical scanning, though advanced, cannot yet digitise written text automatically to a sufficient standard.

This means that there remains some possibility that a few errors, especially doublings and omissions and punctuation errors, may have been introduced accidentally by the typists, who cannot use automatic scanning. However, it is the view of the current authors that rigorous quality control procedures, as used on the CALES project, will ensure that such interventions do not occur in any significant numbers.

Furthermore, the typist/analysts employed on the CALES project were very carefully trained to type the essays exactly as they were written by the students who produced them. Also, possible interference by automatic correction features in the editing software

was removed by using a text editor such as WordPad, or by switching off the AUTOCORRECT options within MS Word. Consequently, once those few putative cases of error on the part of the typist were eliminated, we are still left with a significant amount of mechanical errors in our corpus samples, which is a phenomenon worthy of further study in a larger corpus sample.

Let us now address the question of why there is such a high frequency of mechanical errors, by focusing on punctuation. One reason put forward for the high frequency of punctuation errors in our corpus may be that learners (and perhaps teachers in some cases) may be confused about the rules, despite the rules being quite clear and explicit. This issue was discussed poignantly by Truss (2003) who pointed out that even native speakers are confused about punctuation, perhaps because of insufficient formal instruction in school. One infamous example is the so-called 'market-stall apostrophe', which produces such examples as 'potato's' and '10 kilo's'. Therefore, it may be no surprise that learners of English, when they start to write in English, may be confused also, despite the fact that Malaysian students will have received plenty of instruction in punctuation by the time they have taken their Form 5 examinations.

Indeed, some of this confusion may also be passed on to students from their language teachers, especially if these teachers were not educated in English-speaking countries. It is beyond the scope of this paper to investigate this issue in detail, but finding an explanation for errors in punctuation is a fruitful area for further research and debate. However, we can say that the high frequency of mechanical spelling errors in our data appears to militate against the commonly held belief that spelling errors are largely caused by L1 interference or direct translation.

In our data, there are indeed examples of L1-influenced spelling errors, word coinages and US spellings, lending support to the view that the L1, primarily Malay in this case, may influence the errors made by students in the samples (Botley et al, 2005). Furthermore, we must not ignore the effect of lexical borrowing from English into Malay, despite there being relatively few clear examples in our data. Such borrowings seem to cause many spelling errors in the corpus samples, where English borrowings are given Malay-influenced spellings (e.g. <karier> and <fenomena>).

But despite all this evidence of L1 influence, it remains the case that L1-induced spelling errors, whether phonological or lexical, do not fall into the largest group of cases identified in the data analysed here. This is a finding worthy of note, and one which will benefit from further research with a much larger and more varied data set.

A further factor which the data reveals is the seeming lack of evidence for James' categories of confusibles and dyslexic errors. This is perhaps unsurprising, as much of James' work had a therapeutic focus, looking at writing made by children, some of whom may have been dyslexic. However, the data in this study did not appear to reflect any psychological or cognitive phenomena in the students who wrote the essays. Also, the fact that the essays were written in an EFL context may cloud our view of any possible

cognitive factors in the students' performance. Perhaps with further study of a larger data set, we may be able to arrive at some observations in this area.

Finally, we need to address the extent to which students at Degree and Diploma level exhibit different error frequencies. The picture that emerges from Table 3 above is somewhat surprising. It would appear that Degree students, in most cases make more errors than Diploma students, which seems at first to be somewhat counterintuitive. It should be remembered, however, that the number of errors made by Degree students overall is much higher in any case, and we need to take into account other variables such as the amount of contact hours in English classes experienced by Diploma and Degree students. Furthermore, the amount of data analysed in this study is not enough to allow meaningful comparisons in terms of level of study to be made at this stage.

Now that we have discussed the major findings of this study, we will conclude the paper and provide some overall comments and recommendations.

Conclusion

This paper has had the lofty aim of empirically analysing and classifying spelling errors in a sample of essays written by Malaysian university students. Using Computer-Aided Error Analysis (CEA) techniques, we have managed to identify and classify a large number of spelling errors, and provide a number of observations and discussion points.

From one perspective, some of the findings were surprising, and even somewhat shocking. It appears that Degree-level students make more spelling errors compared to Diploma-level students. It seems that there is a high frequency of quite basic mechanical errors such as punctuation and omission. There does not appear to be any evidence of cognitive impairment leading to any of the errors identified. And finally, although there is plenty of evidence of L1 interference and lexical borrowing in the spelling errors made by students in the sample, these errors appear to be much less frequent than other more mechanical categories. Much of this runs counter to what might be expected.

On another level, however, the findings provide few blindingly new insights into the spelling performance of Malaysian university students. Most teachers and lecturers could have predicted the types of errors found in the CALES data, though they may not have been able to predict the frequencies. However, this study has given empirical support to many common perceptions about spelling performance in Malaysian education institutions, by providing hard statistics, and by providing some helpful new insights.

This study reveals the benefits of using corpus data in two important ways. Firstly, it shows how traditional pen and paper Error Analysis can be updated, empowered and enhanced by the use of large amounts of corpus data. Our analysis revealed a number of findings about spelling errors which were different from James' original analyses of spelling errors, based as they were on analysis of different but authentic data. We were

able to in some cases re-visit some of James' categories, and found many examples of errors which did not fit exactly into those categories.

Secondly, our study has shown that Computer-Aided Error Analysis can reveal a great deal of valuable insights by combining a labour-intensive data input process with the speed and automation of the computer. There appears to be no realistic alternative to this semi-automatic style of analysis at the moment. Even though it is possible to write software to identify spelling errors, it is another thing again to place them into meaningful categories that allow us to understand why learners make them.

Finally, then, this study has shown that there is still a lot of room for improvement in spelling and punctuation among university EFL students, at least in Sarawak and Sabah. Students still seem to be making many basic mechanical errors, as well as some errors influenced by the L1, despite many years of intensive English language instruction.

This would lead us to recommend that punctuation and spelling should be given more emphasis in future curriculum planning in Malaysia, as it is clearly causing difficulties for students, even at the higher education levels. Furthermore, an emphasis on regular reading practice on the part of language learners would be a highly useful way of exposing students to plenty of examples of 'correct' spellings, thereby helping them to unconsciously avoid making spelling errors in their writing practice.

This paper ends on a relatively positive note. In future research, using more data, it will be possible to make more detailed and generalisable statements about spelling errors in Malaysian student writing. It is hoped that findings derived from such research will help us to devise better methods and teaching materials for Malaysian university students, as well as providing a deeper understanding of Malaysian students' interlanguage in mastering the English Language.

References

Altenberg, B. (2002). Using Bilingual Corpus Evidence In Learner Corpus Research. In Granger, S., Hung, J. and Petch-Tyson, S. (Eds.). *Computer Learner Corpora, Second Language Acquisition And Foreign Language Teaching* (pp. 37-54). Amsterdam: John Benjamins.

Arshad Abd. Samad, Fauziah Hassan, Jayakaran Mukundan, Ghazali Kamarudin, Sharifah Zainab Syd Abd. Rahman, Juridah Md. Rashid & Malachi Edwin Vethamani. (2002). *The English Of Malaysian School Students (EMAS) Corpus*. Serdang: Universiti Putra Malaysia.

Barlow, M. (2005). Computer-based Analyses of Learner Language. In Ellis, R. and Barkhuizen, G. (2005). *Analysing Learner Language.* (pp. 351-352). Oxford: Oxford University Press.

Botley, S. P., De Alwis, C., Metom, L., and Izza, I. (2005). *CALES: A Corpus-Based Archive Of Learner English In Sarawak. Final Project Report,* Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Botley, S. P., Metom, L., and Dillah, D. (2007). *A Corpus-Based Archive Of Learner English In Sabah/Sarawak: CALES Phase 2. Final Project Report*, Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Carney, E. (1994). *A Survey Of English Spelling.* London: Routledge.

Dagneaux, E, Denness, S. and Granger, S. (1998). Computer-Aided Error Analysis. System. *An International Journal Of Educational Technology And Applied Linguistics,* 26(2), 163-174.

Granger, S. (Ed.) (1998). *Learner English On Computer*. Harlow: Addison-Wesley Longman.

Granger, S., Hung, J. and Petch-Tyson, S. (Eds.) (2002). *Computer Learner Corpora, Second Language Acquisition And Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S., Dagneaux, E. and Meunier, F. (2002). *The International Corpus Of Learner English Handbook And CD-ROM.* Louvain-la Neuve: Presses Universitaires de Louvain.

Granger, S. and Tribble, C. (1998). Learner Corpus Data In The Foreign Language Classroom: Form-Focused Instruction and Data-Driven Learning. In Granger, S. (Ed.) *Learner English on Compute*r, (pp. 199-209). Harlow: Addison-Wesley Longman.

James, C., Scholfield, P., Garrett, P. and Griffiths, Y. (1993). Welsh Bilinguals' English Spelling: *An Error Analysis. Journal Of Multilingual And Multicultural Development,* 14(4), 287-306.

James, C. (1998). *Errors In Language Learning And Use: Exploring Error Analysis*. Harlow, Essex: Addison-Wesley Longman.

Kaszubski, P. (1998). E*nhancing A Writing Textbook: A National Perspective. In Granger, S. (Ed.) Learner English On Computer,* Harlow: Addison-Wesley Longman, 1998, pp. 172-185.

Knowles, G. and Zuraidah Mohd Don. (2004). Introducing MACLE: The Malaysian Corpus Of Learner English. 1st National Symposium of Corpus Linguistics and Foreign Language Education, South China Normal University, Guangzhou, China, 10–14 October 2004.

Meunier, F. (1998). Computer Tools For The Analysis Of Learner Corpora. In Granger, S. (Ed.) *Learner English On Computer*, (pp. 19-37). Harlow: Addison-Wesley Longman.

Milton, J. (1998). Exploiting L1 And Interlanguage Corpora In The Design Of An Electronic Language Learning And Production Environment. In Granger, S. (Ed.) *Learner English On Computer*, (pp. 186-198). Harlow: Addison-Wesley Longman.

Richards, J. C. and Schmidt, R. (2002). *Longman Dictionary Of Language Teaching And Applied Linguistics.* Third Edition. Harlow: Pearson Education Limited.

Ringböm, H. (1998). Vocabulary Frequencies In Advanced Learner English: A Cross-Linguistic Approach. In Granger, S. (Ed.) Learner English On Computer, (pp. 41-52). Harlow: Addison-Wesley Longman.

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

Selinker, L. (1972). Interlanguage. *International Review Of Applied Linguistics,* 10(3), 209-31.

Truss, L. (2003). *Eats, Shoots And Leaves: The Zero Tolerance Approach To Punctuation*. London: Profile Books Ltd.

Appendix 1

Sample Essay.

1. **uitmskbsjdip0045**:

Money is important in our life. Everyones are working so that they can find the money. Without the money our life is no meant because if we are poor people, no ones want being a friend or take care of us. For poor family it is hard to get the money but the rich person easily waise the money. That a fact of our life. In fact of requirement of the worth, the money become the root of all evil. Many cases occur in our country as the causes of money.

One of the cases occur in our country are thiefs and the robbers. They aim to get the luxuries by thief or rob because it is a short way to find the money. With one operation they can relax then usually the robbers find the bank or the richmens' houses and the large shop so that they can get alot of money as them can than the small shop or ordinary person's houses. Some of the thief injuring victims. This always happen during the festival seasons such as Hari raya Aidilfitri, Chinese New Year or Deepavali Day.

Argument among family member is the second of the causes. For example, if the father is passed away, his worth must be transfer to the children and his wife. This will make a difficult to get the agreement among them. This always happen in our society. because the distributenent of the luxuries are not fair. Also it can cause fight among the family. When bad attitude influence in distributed the worth, the relationship of the family will break.

Also, the money can increase the numbers of the greedy person. This always happen in our nation especially the seller, shopkeeper or the private clinic. They want to collect more the money than the money they might to get. For example the seller cheats in the charge or total amount that the costumer must to pay. Sometimes, they increasing the price of the certain goods such as subsitute goods and compliments goods. If the government does not control the price of this goods, the consumer may be suffered.

Lastly someones can being a killer. As in fact of the money, they able to kill somebody to fullfill their desires or needs. The number of the kidnappers also increase. This case is the easy way to get the luxuries. A richman and well known person always being the aim of this person to get the money.

In a conclusion, money is the root of all evil. The government should play their role to solve this problem.

Appendix 2

List Of Essay Titles Used In The CALES Projects

1. Crime does not pay. How far do you agree with this statement?
2. Most university courses are too theoretical and do not prepare students for the real world. Discuss.
3. Money is the root of all evil.
4. Science and technology have brought more harm than good.
5. The death penalty should be imposed for rapists.
6. It is a great thing to fight for one's country.
7. The price of development is an erosion of our moral values and culture. Discuss this statement.
8. Our lives are increasingly becoming dependent on computers and other sophisticated gadgets. Do you agree?
9. Popular entertainment programmes such as *Akademi Fantasia* or *Malaysian Idol* are just harmless entertainment. Do you agree with this statement?
10. Foreign workers are essential for the rapid development of our country. Discuss.