

**A CORPUS-BASED INVESTIGATION OF THE INTERLANGUAGE OF  
UNIVERSITY STUDENTS IN EAST MALAYSIA**



**RESEARCH MANAGEMENT INSTITUTE  
UNIVERSITI TEKNOLOGI MARA MALAYSIA  
40450 SHAH ALAM,  
SELANGOR, MALAYSIA**

**PREPARED BY:**

**PUAN LILLY METOM  
PROF. MADYA DR. SIMON BOTLEY  
@ FAIZAL HAKIM**

**2011**

## TABLE OF CONTENTS

	PAGE
TITLE PAGE .....	1
LETTER OF SUBMISSION .....	2
RESEARCH TEAM MEMBERS .....	3
TABLE OF CONTENTS .....	4
ACKNOWLEDGEMENTS .....	8
LIST OF TABLES .....	15
LIST OF FIGURES .....	16
ABSTRACT .....	17
CHAPTER 1: INTRODUCTION .....	18
1.1 Research Background and Problem Statement .....	18
1.2 Objectives .....	22
1.3 Significance .....	22
1.4 Scope and Limitations .....	24
1.5 Definition of Terms .....	26
CHAPTER 2: LITERATURE REVIEW .....	28
2.1 How Corpus Linguistics Developed .....	28
2.2 The Rise of the Learner Corpus .....	33
2.3 Learner Corpora in South-East Asia and Malaysia .....	38

Table 1: Current CALES Corpus Composition (including CALES 1) .....	57
Table 2: Age of Students in CALES Corpus .....	58
Table 3: Gender .....	59
Table 4: Religion .....	59
Table 5: Registration Status .....	60
Table 6: Ethnicity .....	60
Table 7: Home Town Origins (Most frequent, selected) .....	62
Table 8: Home States of CALES students .....	63
Table 9: School locations .....	64
Table 10: Medium of Instruction at Primary School .....	65
Table 11: Medium of Instruction at Secondary School .....	65
Table 12: Medium of Instruction at Post-Secondary Level .....	65
Table 13: Medium of Instruction at Diploma Level (UiTM only) .....	66
Table 14: Medium of Instruction at Degree Level .....	66
Table 15: Year and Semester of Study .....	67
Table 16: Institution Attended .....	67
Table 17: Years Learning English .....	68
Table 18: Self-Reported Native Language .....	70
Table 19: Parents' Native Language .....	72
Table 20: Most Frequent Language Spoken at Home .....	72
Table 21: Second Most Frequent Language Spoken At Home .....	73
Table 22: Third Most Frequent Language Spoken at Home.....	74
Table 23: Percentage of Use of Most Frequently Spoken Home Languages .....	74
Table 24: Percentage of Use of Second Most Frequently Spoken Home Languages .....	74
Table 25: Most Proficient Languages of CALES Respondents.....	75
Table 26: SPM English Grades .....	76
Table 27: MUET (Malaysian University English Test) Bands .....	77
Table 28: English Tuition Attendance .....	77
Table 29: Programme and Faculty of Study .....	78
Table 30: Level of Study .....	79
Table 31: Essay Titles .....	79
Table 32: Essay Length .....	80
Table 33: Use of References .....	81
Table 34: Which References? .....	81
Table 35: LOCNESS Sub-corpus, Compared with CALES Sub-corpus.....	87
Table 36: N-grams in CALES Sub-Corpus.....	89
Table 37: N-grams in LOCNESS Sub-corpus.....	92
Table 38: Comparing frequencies of learner-generated n-grams with a native speaker benchmark .....	95
Table 39: Pronoun Errors: Essential Statistics.....	103
Table 40: Pronoun Errors and Target Expressions in Diploma-Level Essays.....	104
Table 41: Target expressions for erroneous 'this' as demonstrative determiner.....	106
Table 42: Target expressions for erroneous 'it'.....	107
Table 43: Target expressions for erroneous 'their'.....	107

## ABSTRACT

This report describes the culmination of the learner corpus project called CALES (Corpus-based Archive of Learner English in Sarawak). The original two phases of the project collected 356,000 words of learner writing in the form of argumentative essays written by students taking English proficiency courses in UiTM's Sarawak Branch Campus (Botley et al, 2005, 2007). This new project has increased this total to over 480,000 words of digital text, and has collected more essays from four different institutions in order to further expand and enrich the corpus.

The project follows the methodological principles laid down by the International Corpus of Learner English (ICLE) project in Belgium (Granger et. al., 2002). The data was digitised and analysed in order to investigate different types of language error. A number of observations were made concerning some prominent error categories in the data, and their pedagogical implications were explored.

It is hoped that these findings will further contribute to our understanding of the way in which Malaysian learners of English actually perform in their writing. Also, it is hoped that the outcomes of this project will form a foundation for a larger-scale understanding of the interlanguage (Selinker, 1972) of Malaysian EFL learners at university level, as well as providing a data resource for future research in this area.



## CHAPTER 1: INTRODUCTION

### 1.1 Research Background and Problem Statement

Educators in EFL (English as a Foreign Language) are all too aware of the errors, or performance features<sup>1</sup> frequently found in writing produced by students of English. However, EFL educators are often unable to make full use of the information revealed by such features in order to help students to improve their language performance. One reason for this is a lack of reliable and clear examples taken from real student texts. Such examples could then be used as a source of reference to help teachers predict the features students may display in their writing and speech, and then do something about them in a systematic and principled manner.

Many EFL educators tend to rely upon their professional experience and linguistic intuitions to predict what kinds of features will be displayed in the writing of a particular non-native-speaker group. For instance, it is widely known that Malaysian learners of English regularly under-use the definite article, and turn non-countable nouns onto countable ones (*a staff*, rather than *a member of staff*).

Errors such as these may be traced back to the L1 which in most cases in Malaysia is Bahasa Melayu, a language which does not have a system of definite and indefinite articles, and in which the notion of countability is somewhat different to that in English (see Botley, Haykal and Monaliza, 2005 for a recent discussion of this issue).

---

<sup>1</sup> Here, we prefer the term 'performance features', because common terms such as 'errors' or 'mistakes' can be considered judgemental and prescriptive. Furthermore, see the section on definition of terms below.