CALES: A CORPUS-BASED ARCHIVE OF LEARNER ENGLISH IN SARAWAK

INSTITUTE OF RESEARCH, DEVELOPMENT
AND COMMERCIALISATION,
UNIVERSITI TEKNOLOGI MARA,
40450 SHAH ALAM,
SELANGOR, MALAYSIA

PREPARED BY:

PROF. MADYA DR. SIMON BOTLEY
@ FAIZAL HAKIM (KETUA)
MR. CAESAR DE ALWIS
PUAN LILLY METOM
CIK. ISMA IZZA MOHD. ESA

15 MARCH 2005

15 March, 2005

Prof. Dr. Azni Ahmad
Assistant Vice-Chancellor (Research)
Institute of Research, Development and Commercialisation
Universiti Teknologi MARA
40450 Shah Alam
Selangor
Malaysia

Dear Prof. Dr. Azni,

REF: FINAL REPORT FOR PROJECT: "CALES: A CORPUS-BASED ARCHIVE OF LEARNER ENGLISH IN SARAWAK"

With reference to the above, we hereby submit two (2) copies of the final report of this project, as required by URDC regulations. We sincerely hope that it meets with your approval.

Thank you.

Yours faithfully

Prof. Madya Dr. Simon Botley @ Faizal Hakim
Project Head

TABLE OF CONTENTS

	PAGE
TITLE PAGE	1
LETTER OF SUBMISSION	2
RESEARCH TEAM MEMBERS	3
ACKNOWLEDGEMENTS	7
LIST OF TABLES	10
LIST OF FIGURES	11
LIST OF CHARTS	11
ABSTRACT	12
CHAPTER 1: INTRODUCTION	13
1.1 Research Background and Problem Statement	13
1.2 Objectives	14
1.3 Significance	15
1.4 Scope and Limitations	16
1.5 Definition of Terms	17
CHAPTER 2:LITERATURE REVIEW	19
2.1 How Corpus Linguistics Developed	19
2.2 The Rise of the Learner Corpus	24
CHAPTER 3:METHODOLOGY	28
3.1 Set-Up Phase	29
3.2 Data Collection	29
3.3 Annotation and Analysis	31
3.4 Quality Control and Indexing	34
3.5 Final Checks and Release	35

ABSTRACT

This report describes a new learner corpus project called CALES (Corpus-based Archive of Learner English in Sarawak). The project has collected 89,000 words of learner writing in the form of argumentative essays written by students taking English proficiency courses in UiTM's Sarawak Branch Campus.

The project follows the methodological principles laid down by the International Corpus of Learner English (ICLE) project in Belgium (Granger et. al., 2002), and the data has been analysed in order to classify different types of error found in the corpus. A number of observations have been made concerning the most frequent errors in the data, as well as some correlations between the errors and the social, educational and linguistic attributes of the learners who produced the essays.

It is hoped that these findings will contribute a great deal 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 corpus-building enterprise in the future.

CHAPTER 1: INTRODUCTION

1.1 Research Background and Problem Statement

Teachers of English as a Second Language (ESL) are generally fully aware of the mistakes or errors that students make when attempting to write in English. However despite this awareness, it can be said that not all ESL educators can or do make full use of those errors in order to help students to avoid making them in the first place. One reason for this is a lack of reliable and permanent data on features of learner performance in English. Such data, if it was available, could be used as a reference to help ESL educators to predict what kind of errors students make, and to enable educators to do something about the errors in a systematic and principled manner.

At the moment, most ESL educators rely upon their professional experience or linguistic intuition to predict what kinds of errors in the L2[1] will be made by a particular L1[2] group. For instance, it is widely known among Malaysian ESL educators that Malaysian learners of English regularly over-use the definite article (*we must work for the money* versus *we must work for money*), 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

[1] Second or target language
[2] First or native language