Department of Computer Science,
University of Otago

Technical Report OUCS-2004-04

Analysis Of Errors Made By Learners Of Maori In An Introductory University Course

Authors:

Lorene Earnshaw, Stewart Fleming, Alistair Knott
Department of Computer Science, University of Otago

Victoria Weatherall
School of Maori, Pacific and Indigenous Studies, University of Otago

Status: Accepted for publication in the Journal of Maori and Pacific Development, to appear in September/October 2004

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/trseries/

Analysis of Errors in the Writing of First Year University Students of Maori

Lorene Earnshaw, Stewart Fleming
Software Engineering Group, Department of Computer Science
lorene@cs.otago.ac.nz stf@cs.otago.ac.nz

Victoria Weatherall
School of Maori, Pacific and Indigenous Studies
Victoria.weatherall@stonebow.otago.ac.nz

Alistair Knott
Artificial Intelligence Group, Department of Computer Science
alik@cs.otago.ac.nz

University of Otago, P O Box 50, DUNEDIN, New Zealand.

Abstract

This paper presents a survey of the range of grammatical and lexical errors made in written Maori by University students taking an introductory course in Maori language. We begin by introducing, and discussing the motivation for, an error classification system which accommodates different classes of error. We then provide an analysis of errors in three different types of student writing: homework assignments, impromptu tests and examination transcripts. We conclude with some remarks about the patterns of error-which we found.

Acknowledgements

Lorene Earnshaw is supported by funding from the Foundation for Science, Research and Technology (FRST) under contract no. UOOX0307. Victoria Weatherall and Alistair Knott are supported by FRST grant UOOX0209. We are grateful to our colleague Richard O'Keefe for his comments on early drafts of this paper.

Background: Motivation for the study

When teaching a foreign language, it is important to have in mind a model of the kinds of error to which students are prone at any point in their course. What a teacher should do with this model is an issue which has generated a certain amount of controversy. For instance, it has been argued that explicit error correction strategies are unhelpful as they discourage students from using constructions which they are uncertain about (Truscott, 1996). Others argue that error correction can be beneficial when it is given in a form which students can readily understand (see, for example, Ferris, 1999; Lee, 1997). However the issue of explicit error correction is resolved, it seems clear that the teacher needs to know about the kinds of error which are commonly made by students (Corder, 1967). Without this knowledge, it is hard to decide how quickly to proceed from one topic to another, or which types of interaction to concentrate on in the classroom.

In this paper, we present an analysis of the kinds of written error made by University students taking an introductory course in Maori language. Our motivation is twofold. First, we want to provide the teacher involved in the course, and (with some caveats),

other teachers of introductory courses in Maori, with some quantitative information about the patterns of error. Secondly, and more specifically, our study is intended to play a role in the design and evaluation of two computer-aided language learning (CALL) systems which we are building to complement the course: one developed by a Software Engineering group, and one developed by a Computational Linguistics group.. We want to ensure that these systems are designed to target the errors which are most commonly made. Moreover, we aim to incorporate into our evaluation of the CALL systems a comparison of errors made by students who have had experience with one CALL system or the other and students who have experienced only face-to-face teaching. Our focus here is, however, on the analysis of errors itself rather than on the CALL systems to which reference has been made, our main purpose being to present what we hope will be a useful resource for teachers involved in introductory courses in Maori language.

We begin by surveying some existing work on error analyses in Maori. We then present the classification system we developed to accommodate the types of error we encountered. This classificatory system draws on earlier approaches to classification, and extends them in some respects. The analyses of several different types of student work using this classification system are then presented, along with some commentary on the reliability of the classification. Finally, we draw some tentative conclusions for teachers of courses similar to the one under consideration here.

It is important to note here that although earlier work on the analysis of errors made by learners of Maori focused on involved school-age children (particularly those enrolled in the kura kaupapa system), our student population is different, being made up of University students, a mixture of Maori and non-Maori, most of whom have very little background in the language.

Proficiency development and error analysis relating to learners of Maori: A review of some existing research

The last few years have seen the emergence of research on language proficiency development and error analysis in the case of learners of Maori. Johnson and Rolleston (2001) report on the results of pre- and post-course Maori language proficiency testing in the case of teachers attending an in-service development course at the University of Waikato. Crombie, Houia and Reedy (2000) comment on the development and trialling of a pilot proficiency test designed for use with young learners of Maori (particularly those in Year 5 of Maori-medium schooling). Both of these studies, particularly the second, are relevant to a study by Houia (2002) which analyses typical errors produced by the Year 5 students who took part in the proficiency test trials referred to by Crombie, Houia and Reedy. Houia's research presents the most detailed study of Maori language errors to date, and was a particularly valuable source of reference for the development of the error classification system presented here.

At the top level, Houia's classification system distinguishes among errors in relation to whether they involve omission, addition, selection or ordering. Omission errors involve the exclusion of elements; addition errors involve the inclusion of extra elements; selection errors involve choice of the wrong element. In the case of ordering errors, all of the obligatory elements are present but are wrongly ordered. Houia's classification system also distinguishes among errors in relation to another dimension: whether they are grammatical, lexical or semantic in nature. In addition, it provides, wherever possible, a detailed analysis of each error in terms of the precise nature of the omission, addition, selection or ordering problem encountered.

Student Profiles

The students whose errors we investigated were enrolled in an introductory Maori language course - MAOR 110 - at Te Tumu (the School of Maori, Pacific and Indigenous Studies) at the University of Otago. This paper assumes no prior knowledge of the Maori language. The objectives of the course are for students to learn some basic sentence structure patterns of Maori, to develop confidence in using these patterns in both writing and in speech in context, to develop basic conversational fluency, and accurate pronunciation. The topics cover the first three chapters of *Te Kākano* (Moorfield, 2001), with an audio CD for listening exercises.

Students who enrol in MAOR 110 come from a range of different discipline backgrounds. Some are pursuing a major in Maori Studies; others are, for example, pursuing studies relating to professions in which a knowledge of Maori will be beneficial. Although the course under consideration here is a 100-level paper, it attracts a mixture of first year students and more advanced students, school leavers, mature students and staff. The paper is designed for students with no prior knowledge of the Maori language so most of the students are first time Maori language learners. Even so, there is generally a small number of students who have prior knowledge of the Maori language. Most of these students have not, however, learnt the language in a formal environment.

Course assessment involves a mixture of assignment types (see *Table 1* below):

Table 1: *Assessment Structure for MAOR110*

Internal Assessment:		60%
Listening/Oral	30%	
Listening	20%	
Written	10%	
Final Exam (Written) 2 hours		40%

Listening and oral assessments are based on individual tutorials. Students are expected to attend these tutorials at pre-arranged times and to undertake assessments based on listening exercises (with associated questions) included the *Te Kakano* CDs Written assessments are given to students to complete in their own time. Students must return the assessments within a certain time frame (normally one week after the assessments are handed out). Other tests can be introduced into the course without prior warning, and a final examination is completed by students at the end of the course.

Classification of errors

Our work is motivated by the need to consider the types of error made by students learning te reo Maori. One reason for doing this is to understand how additional teaching support could be provided via provision of computer-aided tutorial systems. Another reason is to develop more sophisticated linguistic models to be used in the construction of language translation tools. To suit research projects targeted at these different areas, we had to develop a classification scheme that was flexible enough to record information about the types of error that were commonly made.

The error classification system

Since the activities of our two research groups (Computational Linguistics and Software Engineering) are differently motivated, each group initially adopted a different approach to the construction of the classification system. The Computational Linguistics group proceeded in a top-down manner, defining broad categories first. The Software Engineering group pursued a bottom-up approach, beginning by identifying specific errors in each text, and developing and adapting their classificatory system in terms of best fit as new errors were encountered and compared with errors that had already been identified. This approach allowed new levels of detail to be incorporated into the classification progressively, allowing for both the recording of new errors as they were encountered and the provision of more detail in relation to errors already recorded.

By adopting a common coding scheme, the two groups were able to evolve separately at the same time as maintaining an exchange of data. Thus, we were able to achieve a fluid, flexible state of affairs regarding the coding of errors for different purposes. For example, although the Software Engineering Group added more detail about errors involving macrons than was required by the Computational Linguistics group, this did not affect the exchange of data.

We began our error classification by examining the research of Houia (2002), where errors are roughly grouped into categories according to whether they involve problems associated with selection, addition, omission or ordering. However, we found that these broad categories did not provide enough precision for our purposes in developing computer support for te reo Maori learners. It suited our purpose better for these to become sub-categories and for the main categories of error to be divided into grammatical and content/dialogue errors. In fact, we found that we frequently had to rearrange our classification system to accommodate errors in the dataset under consideration that had not been encountered in the analysis of previous datasets. This suggested to us the need for a more flexible arrangement of categories and sub-categories, with automated (where possible) exchange of data between them. Regarding the form of the classification, the guiding principles were that the system should be:

- quick and easy to use;
- describable in terms of a simple, standardised procedure;

- flexible enough to allow for the introduction of new types of error as they were encountered and for the revision of categories to accommodate further detail;

- organised in such a way as to capture detail about specific errors as well as information about general types of error.

This last point is critical for the effectiveness of our work as it allows for a situation in which the actual error analysis and any subsequent software development based on that analysis can be carried out independently of one another, with both groups (error analysts and software developers) being able to record and/or recover all of the information required at different stages throughout the history of the project.

Coding errors

In this section, we discuss each of the categories in turn, referring to the different aspects of the classification shown in *Tables 2, 3* and *4*. The top-level classification distinguished between closed-class word errors, open-class word errors and dialogue errors. We included two extra categories: word ordering, and a catch-all category 'Wrong'.

Closed class errors (Table 2) are errors involving those word classes (such as pronouns, tense/aspect markers, determiners and prepositions) that include a specific, easily identifiable number of items. Kinship terms are normally regarded as part of the 'noun' open class, but they are particularly important in te reo Maori, and do form a closed subclass of nouns. In particular, they are taught as if they were closed class words.

Open class errors (Table 3) refer either to errors involving open class words, such as nouns and verbs, or to errors which could involve words of any class, such as spelling errors. Open-class errors are divided into word selection errors (choosing the wrong word), word addition errors, word omission errors and spelling errors. We also distinguish two extra classes of error - macron errors (which we have treated separately from other kinds of spelling errors) and 'noun phrase' errors, which are errors relating to the selection, addition or omission of a whole noun phrase. Macron errors are treated separately from other spelling errors, not on prior linguistic grounds, but because the researchers wanted to know how many errors could be associated to words involving macrons.

Dialogue errors (Table 4) suggest that the student did not understand the question, or did not understand the dialogue conventions of te reo Maori; these are therefore treated as comprehension errors rather than performance errors. We did not classify errors in dialogue structure or rhetorical structure, because the dialogue structure is controlled and short answers expected. "Did not attempt the question" is included as a Dialogue Error. This is because there can be a number of different reasons for failing to answer a question, including lack of understanding of the question itself .

Ordering Errors belong to the word ordering category. Since the range of syntactic structures used in MAOR110 is fairly limited, this class needs no subclasses.

The final top-level class, *"Wrong"*, is used for errors that do not fit well into any other class. In our analysis of source texts, this classification was used for answers which had more than three distinct errors. This class of error was used when there were more than three errors present in a sentence, or when the sentence was 'irreparably wrong'.

An error classification system designed for a more advanced paper might, however, require syntactic sub-classification. It might, for example, be possible to distinguish between structural errors at the phrase level and structural errors at the sentence level. However, the first kind (*e.g.* some noun phrase errors) tend to show up as "addition" or "omission" errors in the word error classes. Morphological errors (such as errors involving a passive suffix) were not recorded as they are not directly relevant to the content of MAOR110 and were therefore not tested. They would, however, be included in the analysis of errors in the case of more advanced classes.

Table 2: *Closed class error classification*

Sub Category	Type	Error Type	Sub Type	Code	Example
Pronouns	Personal Pronouns	selection	Used Possessive Pronoun	PPSP	Kei te hiainu taku.
			Mimicking	PPSM	Q Nō hea koe? A. Nō Rotorua koe.
			Used instead of determiner	PPDET	Ko Anaru ia kuri
			Used Wrong One	PPSW	Q Nō hea koe? A. Nō Rotorua ia.
		addition	Addition	PPA	Ko John taku matua rāua ko Miriam taku whaea.
		omission	Omission	PPO	Ko John Ko Jame aku tuākana
	Possessive Pronouns	selection	Used Personal Pronoun	POPSP	Ko Hone ia ingoa.
			Mimicking	POPSM	Q. Ko wai tō matua? A. Ko James tō matua.
			Used plural instead of singular	POPP	Q. Ko wai tō matua? A. Ko James aku matua.
			Used singular instead of plural	POPS	Q Ko wai ō mātua. A. Ko James rāua ko Viv taku mātua
			Used instead of determiner	POPDET	Q E hia ngā pene? A. E rua aku pene.
			Used the wrong one	POPW	Q Ko wai tō whaea. A. Ko Viv tana whaea.
		addition	Addition	POPA	Ko John taku hoa taku.
		omission	Omission	POPO	Q Ko wai tō whaea. A. Ko Viv whaea.
Starters	TAMs	selection	Selection	TAMS	Q. Nō hea koe? A. Kei Ōtepoti ahau.
		addition	Addition	TAMA	Kei te tātahi rātou.
		omission	Omission	TAMO	Q. Nō hea koe? A. Ōtepoti ahau.
	Counting TAMs	selection	Used e instead of toko	CSE	e rua aku tuākana
			Used toko instead of e	CST	Tokorua ngā pene
			Used another TAM	CSO	Ko rua ngā kūri
		addition	Used in front of tekau	CTE	E tekau ngā pukapuka
			Used in front of tahi	CTA	Toko tahi te tangata
		omission	e	COE	rua ngā pene
			toko	COT	rua aku tuākana
	Adressee Term	selection	Selection	ATS	Kia ora a hoa.
		addition	Addition	ATA	Kia ora e Henare.
		omission	Omission	ATOM	Kia ora hoa.
	Determiners	selection	Used plural instead of singular	DETSP	Kotahi ngā pene
			Used singular instead of plural	DETSS	E waru te pene
			Used instead of possessive	DETPOP	Q: E hia ō pene? A: E rua ngā pene.
			Used instead of personal	DETPP	Q: E hia ana pene? A: E rua ngā pene.
		addition	Addition	DETA	te tekau
		omission	Omission	DETO	E waru pene
Parts of Speech	Personal Articles	selection	Selection	PAS	Kei te haere ko John ki te toa.
			Used instead of object Marker	POM	Kei runga a te tepu.
		addition	Addition	PAA	Kei te haere ko a John ki te toa.
		omission	Omission	PAO	Kei te haere John ki te toa.
	Object Marker "i"	selection	Selection	OMS	Kei runga o te tepu.
			Used instead of Personal Article	OMP	Kei te haere i John.
		addition	Addition	OMA	Kei runga i te pene i te tepu.
		omission	Omission	OMO	Kei runga te tepu.
	Preposition	selection	Selection	PS	Kei te haere ahau i Ākarana/ Kei raro ki te tepu
		addition	Addition	PA	Kei te haere ki ahau ki Ākarana.
		omission	Omission	PO	Kei te haere au Ākarana/Kei raro te tepu
	Interogatives	selection	Selection	IS	using hia instead of hea
		addition	Addition	IA	Kei hea Ākarana taku kāinga.
		omission	Omission	IO	Kei tō kāinga?
	Possessive Articles	selection	Used o instead of a	AO	Ko John te tamaiti o Harry.
			Used a instead of o	OA	Ko Harry te matua a John.
	Locatives	selection	Selection	LOS	Kei runga. Should be Kei raro.
		addition	Addition	LOA	Kei runga te pene i runga i te tepu.
		omission	Omission	LOO	Kei te tepu. Should be Kei runga i te tepu.
Nouns	Numbers	selection	Selection	NS	E iwa rau. (800)
		addition	Addition	NA	E iwa rau iwa. (900)
		omission	Omission	NO	E iwa rau e whā tekau mā. (941)
	Kinship	selection	Used plural instead of singular	KSS	mātua instead of matua
			Used singular instead of plural	KSP	matua instead of mātua
			Used Wrong One	KSW	tuākana instead of tungāne
		addition	Addition	KA	Ko tuakana ko John taku tuakana.
		omission	Omission	KO	Ko John taku.

Table 3: *Open class error classification*

Macrons	omission	Omitted a Macron	MACO	raua
	selection	Macron in wrong place	MACW	Otēpoti
	addition	Added a macron	MACA	whānāu
Noun Phrases	selection	Used plural instead of singular	NPSP	Ko James aku mātua
		Used singular instead of plural	NPSS	Ko James rāua ko Viv taku matua
	addition	Addition	NPA	Ko taku hoa ko James taku hoa
	omission	Omission	NPO	Kei te haere ki te toa.
Wrong Word Used			WWU	Used tūru for table
Spelling			SPEL	Mistakes with nouns, verbs, adjectives, etc
Omission			OMIS	
Addition			ADD	Ko taku tuakana ko John taku tuakana.

Table 4: *Dialogue Error Classification*

Short Answer Format	SAF	Ko John.(Ko John taku tuakana)
Answering in a different format	FORM	Q: Nō hea koe? A: Kei Ōtepoti au.
Wrong Answer	WA	Q: Nō hea koe? A: Kei Ōtepoti taku kāinga.
Used an English word instead of Māori	ENG	Ko John taku brother.
Did not attempt the question	DNA	
Partial Answer	PART	Q: Ko wai ō mātua. A: Ko Sue taku māmā.

Analyses of student texts

The classification system was developed from the in-depth analysis of written assessments that were undertaken by students as a formal part of the studies of MAOR110. In this section, we report on the method of analysis that was used to initially develop the classification and on our method of ensuring reliability of the classification of these errors by the analysts.

Choice of student texts

Our main aim was to choose texts which were representative of the written work which our students had to produce. We have not attempted to generalise beyond the specific course which we are considering. This means that the results are inevitably skewed towards the particular exercises which students undertake on this course.

There are three types of data available: Assignments, Impromptu Tests and the Final Exam. Assignments were exercises that the student took home and completed before a due date. The impromptu test was unannounced and the students were not allowed access to any other written material while completing it. The final exam was carried out at the end of the semester, covering all the topics that had been taught up to that point. The final exam contained elements that were not present in the previous two datasets. As we did not track the work of individual students, we were unable to compare the performance of different individuals across the three data sets.

We found that we had to make some more detailed decisions about coding of errors as we were coding the source texts. The context of the questions and answers had to be taken into account when analysing the errors. We found cases where, while an answer was grammatically correct it did not answer the specific question being asked. We also found cases where a student provided a more complex response than the one that was anticipated. Note that these two categories of answer are the ones that are likely to be most difficult to recognise in a computer-supported learning environment.

Data entry and normalisation

Assessment papers were photocopied from the original scripts submitted by students. Any material that might identify the student was redacted. Each script was allocated a unique identifier to allow tracking of errors back to source materials. Each script was coded by an analyst working with the classification. Where an error was found, it was letter coded as indicated in *Tables 2, 3* and *4*.

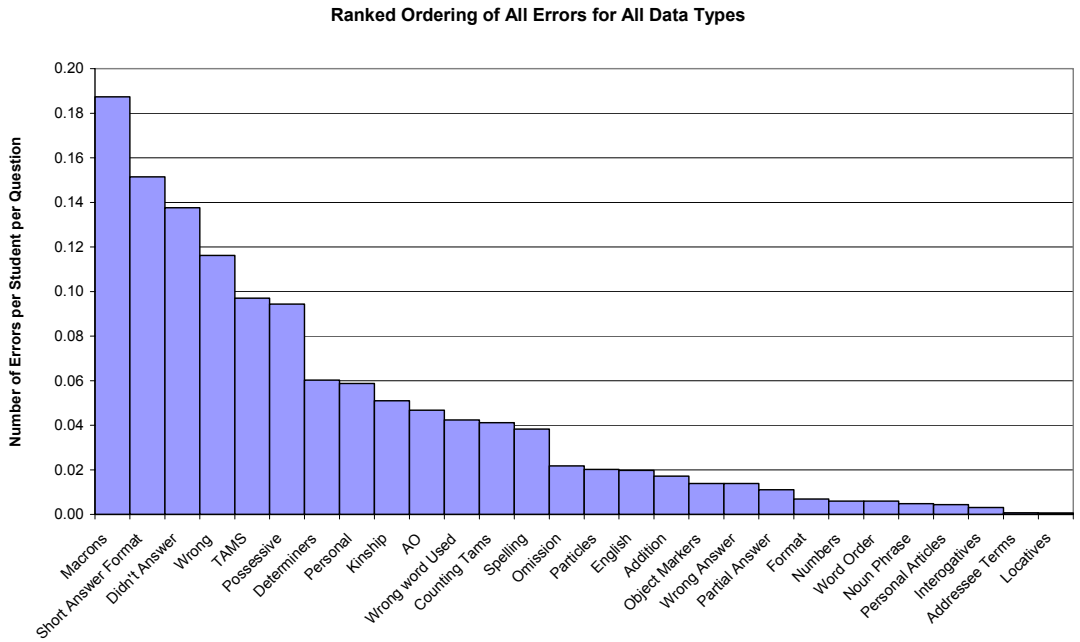
Analysed scripts were collected from the analyst and the classification data entered, tabulated for each individual question. Up to three errors could be recorded for each sub-question. If there were more than three distinct errors for a question, the error type W (Wrong) was used.

The error quantities were normalised for each question by dividing by the number of participants and then by the number of sub-questions available in that particular set. This resulting number is the average number of errors per student per question for each error type.

The most common errors overall

The most common error category resulted from errors involving macrons (see Figure 1), accounting for 15% of the overall total. Errors involving short answer format made up 12% of the total. Half of all the errors recorded are from the top four categories, and approximately 80% are made up from the top ten categories: macrons, short answer format, did not answer, wrong, tense/aspect markers, possessive pronouns, determiners, personal pronouns, kinship and ā/ō.

Figure 1: *Overall Errors*



Macrons

Macron errors were broken down into the following:

- Omission of macron - 92%;
- Addition of macron - 6%; and
- Placement of the macron in the incorrect place - 2%.

By recording data about the error types at the time of coding, we are able to add more detail to assist with further analysis. For example, in *Table 5*, we have collated a list of the most common words associated with macron errors. The word "ma" occurred 77 times (19.4%). By further analysis of this data, we can break down the distribution of errors for a particular word. This ability to view the errors in different ways makes it possible to ask deeper questions about the underlying data and potentially can allow for more targeted teaching and learning activities.

Table 5: *Words associated with macron error*

Word	Occurrence	Word	Occurrence
Mā	77	Whā	45
Pāhi	43	Tungāne	31
Hāpāhi	28	Nō	21
Rā	15	Tēnā	15
Ōtepoti	11	Rāua	11
Wahine	11	Kāore	9
Rākau	9	Kī	8
Rātou	8	Tāone	6
Kāinga	5	Ināianei	4
Tāmaki-makau-rau	4	hāwhe	4
meneti	4	Pāpā	3
Whakatāne	3	Whānau	3
tūru	2	āe	2
mātakitaki	2	tamāhine	2
a	2	tama	2
kāpata	2	haere	2
Ākarana	1	ēnei	1
kara	1		

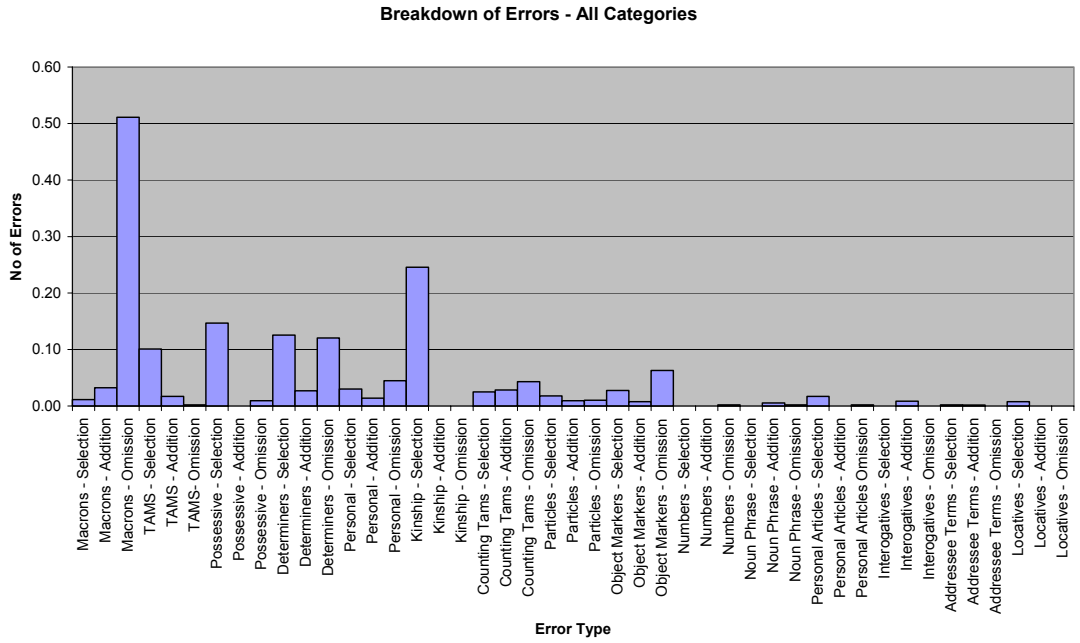
Selection, omission and addition errors

It is also interesting to look at the distribution of error types in terms of the categories of selection, addition and omission (see *Figure 2*). There are several points which emerge from this view of the data. Firstly, errors with macrons are overwhelmingly errors of omission, something that is consistent with the data in *Figure 2* which indicate that the words most frequently misspelled are those that have macrons. Secondly, in the case of closed class words, determiners and object markers are often omitted, while selection errors frequently occur with possessives, tense/aspect markers, determiners and kinship terms.

Most of these findings are consistent with the intuitions of the course teacher. It is unsurprising that omissions frequently occur with object markers and macrons since these do not feature in English. It is somewhat surprising there are not more tense and

aspect marker omissions in general given that these are also not found in English. We assume this is because students can readily understand the rule that most sentences require a tense and aspect marker.

Figure 2: *Breakdown of Errors into sub-categories*

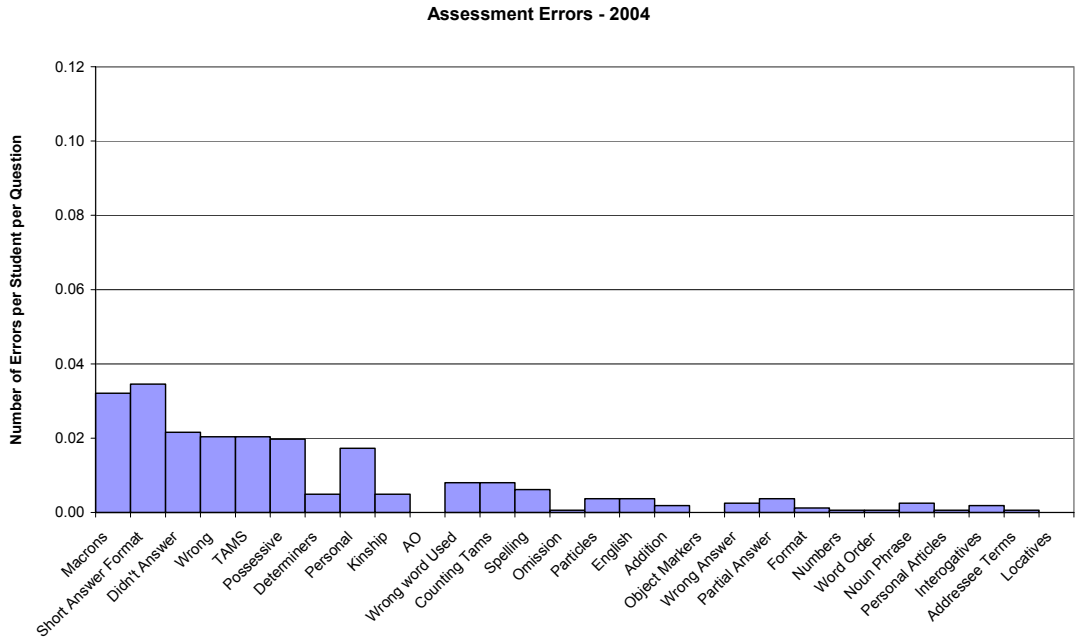


Differences among assessment types

Since the datasets that we use stem from three different sources, we investigated ~~if~~ whether there were any characteristic differences among the assessments. We would expect some variation, since not all assessments tested the same material, nor did they test it in the same way. We plotted three separate charts – assessments (see *Figure 3*), impromptu test (see *Figure 4*) and the final examinations (see *Figure 5*) – using the same categories on the X-axis as in *Figure 2*, and the same scale for the Y-axis on all three. This allows the data in each chart to be compared directly.

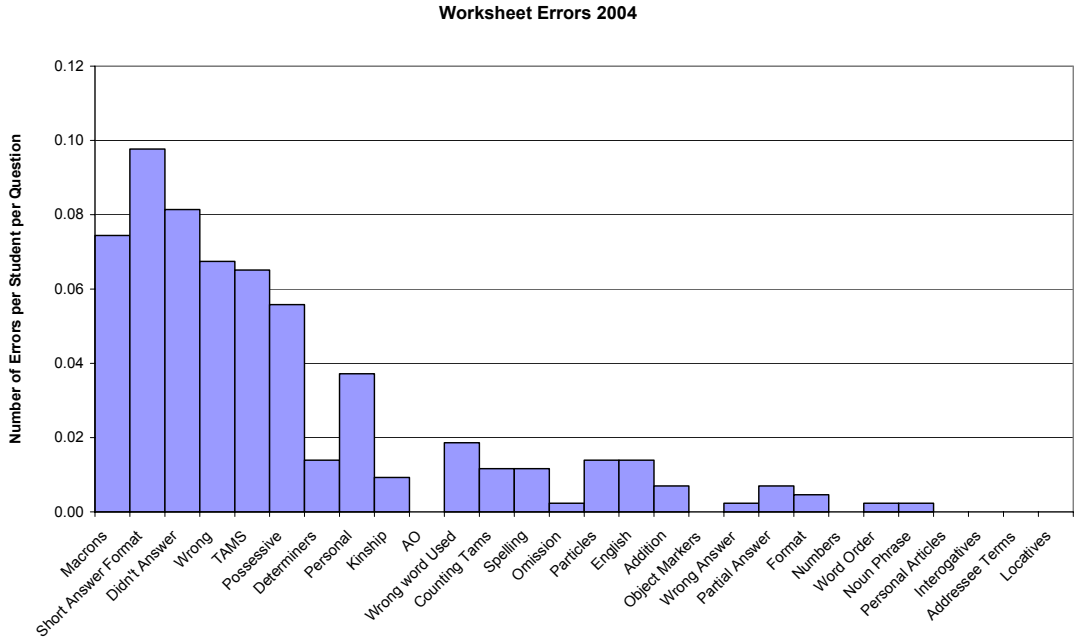
The different datasets we have gathered reflect different sets of questions, which were given to the students at different points in their course. The exams were given at the end of the course, while the assessments and impromptu tests were given at various times throughout the course. As there were elements in some assessments that were not assessed in the others, some error types have zero occurrence in those assessments. It is not possible to address this unequal opportunity for errors to arise in each of the datasets through a reliable normalisation process.

Figure 3: *Assessment Results.*



The data in Figure 3 were collated from the normal coursework-related assessments given to students at various times during the delivery of the paper. For our purposes, we were only interested in the classification of errors rather than the comparison of distributions of error types across assessments.

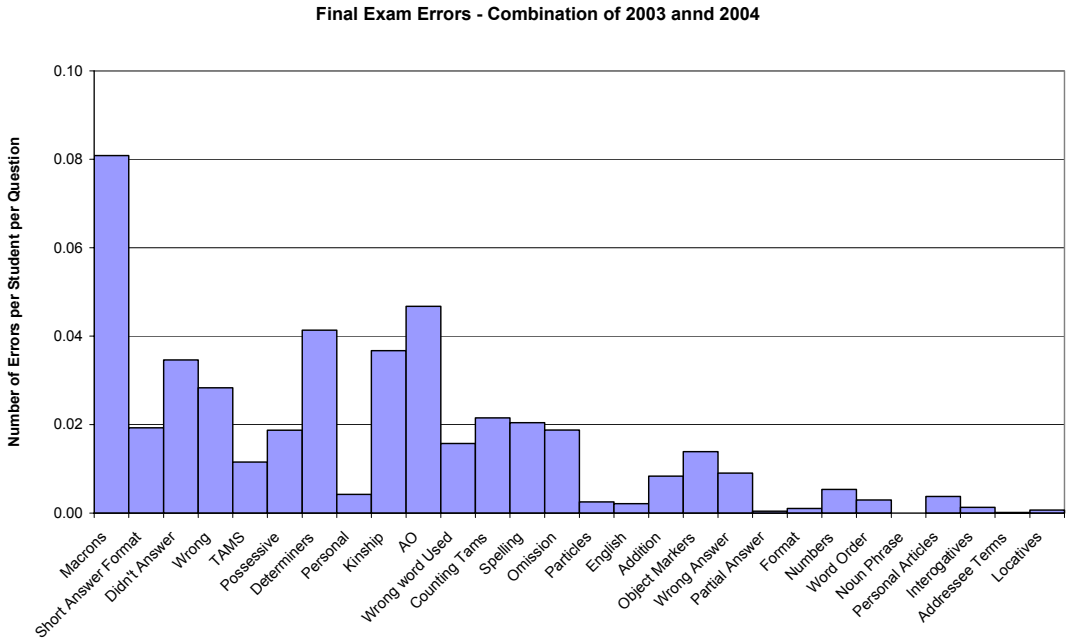
Figure 4: *Impromptu Test Results.*



Data shown in Figure 4 were collated from the test that was given without warning and for which students could not refer to any additional material. Again, we present only the distribution of errors.

Analysis Of Errors

Figure 5: *Final Exam Results.*



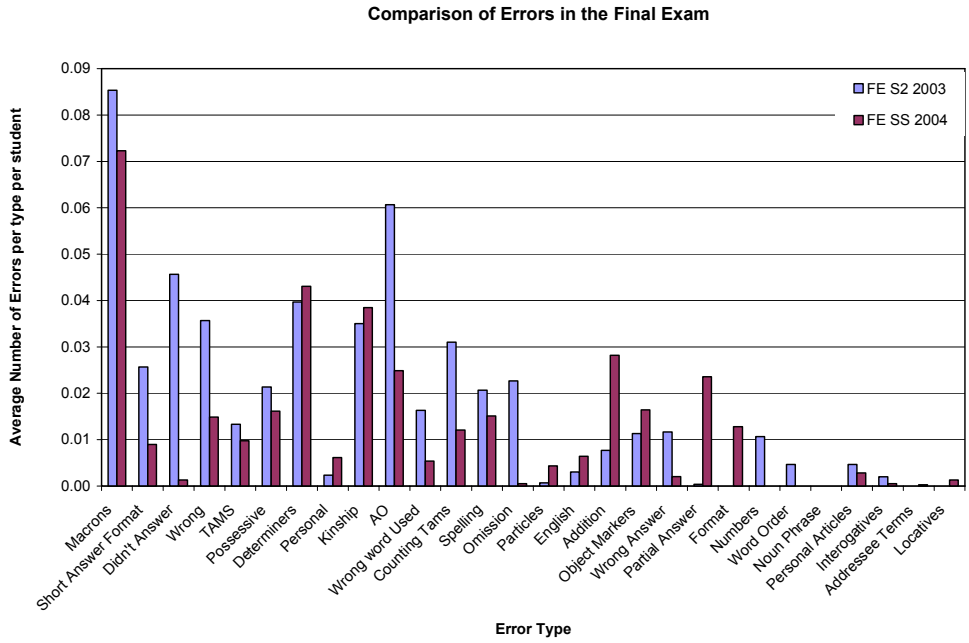
Data in Figure 5 was collected from two final examinations. In these results the error type ā/ō appears, where it does not in the results shown for the assessment and the impromptu test. Students were not given any questions relating to this distinction until the final exam. The absence of this error type in the other two datasets reflects the lack of opportunity for this error to occur.

Differences between classes

Within our datasets, we coded examination scripts for two different final examination papers (see *Figure 6* below). One set of scripts came from the final examination at the end of normal delivery of the paper during semester and the other came from the Summer School version of the same paper.

The ability to record data to track distribution of error types in the assessments across subsequent offerings of a course can be valuable to the teachers of that course. It allows them to determine if variances in teaching technique are having any effect, or to determine long-term reliability of the error classification.

Figure 6: *Comparison of the Final Exams in terms of errors*



Coding process reliability

Since our classification was developed with the notion that separate researchers would be using it for different purposes, we had to consider the protocol (*c.f.* Carletta, 1996) that would be necessary to ensure reliability of the coding process between studies. We tested inter-rater reliability of the coding process by defining a procedure that explained the error categories, provided examples of types of error and how they might be categorised. We used Cohen's kappa statistic (Cohen, 1960) to measure the degree of reliability between the assessments made by different markers.

Procedure

Source material for this exercise was a final examination paper. Six questions were selected from this paper. Each question consisted of a varying number of sub-questions. Data was recorded by two markers (A and B) for each question in the paper. Any personal information that might identify a student was removed from the source material prior to analysis and a separate unique identifier was assigned to each paper for tracking purposes.

Ratings were grouped by question *i.e.* all ratings for sub-questions were tabulated on one sheet. The contingency tables for each question were assembled to cross-tabulate the classifications provided by the two analysts. The overall contingency table was assembled by repeating this process across all questions. Manual and automated error checks were conducted at this point to ensure that all classifications had been entered correctly and that the contingency tables summarised this data correctly.

From the raw data, a contingency table with 2 classification factors (Analyst B against Analyst A) was assembled for each question and an overall contingency table was assembled to aggregate data for all questions. We calculated the kappa statistic for

each table. The kappa statistics overall, and those derived from the individual questions are shown and their implications discussed below.

Results

The following results (see *Table 6*) were computed for the kappa statistic, one for each of the individual question-specific datasets and one for the overall aggregated dataset.

Table 6: *Kappa statistics for the individual questions and overall assessment.*

Data set	Kappa
Q2	0.83
Q3	0.87
Q7	0.97
Q8	0.91
Q9	0.89
Q11	0.82
Overall	**0.90**

There is a high level of agreement between the assessments made by the different analysts, both overall and for the individual questions. The overall agreement is the most significant statistic as it takes into account all opportunities for errors to be classified.

The kappa statistic for the datasets for individual questions is useful as a broad indicator of agreement. However, more useful information about variances between analysts can be found by examining the differences for individual categories. This sort of examination can uncover any misunderstandings of the rating procedure or the classification system, or bias between markers.

Example: In Question 2 of our source material, there are 32 cases where analyst A has classified an error as WWU (Wrong Word Used). In 29 of these cases, analyst B agrees, but has made a different assessment in the 3 other cases. These variations show up in the contingency table as shown in Table 7.

Table 7: *Variations in classifications applied by two individual raters.*

Rater B classification	Rater A classification (WWU)
KSW (Kinship Wrong)	1
NE (No Error)	1
NS (Number Selection)	1
WWU (Wrong Word Used)	29

From knowledge about how the classification was developed, we can start to make more detailed assessments (although slightly more subjective) of the variances.

1. *Kinship Wrong* is a more subtle diagnosis than *Wrong Word Used* and *Number Selection* is an additional distinction;

2. Conversely, *Wrong* answer is a less subtle characterisation.

3. The case where one of the raters marked *No Error* may indicate a variation in the strictness of marking. It may, however, simply reflect the fact that one marker simply missed marking an error.

It is worth noting that this kind of audit trail is enabled by the way the data has been collated and organized. First, we form a pair of classifications – those given by the respective analysts $\{class_A, class_B\}$. Then within the data collated for the question, we search for the classification pair. These are recorded against the unique ID for the script. From the ID, question and sub-question, we can go back to the original script and diagnose why the variation has occurred.

This form of audit activity is extremely useful in improving the procedure that is to be followed by analysts. By giving examples of how the classification is to be used, we can define the procedure in sufficient detail as to minimize variance. We will also uncover cases where we would be willing to accept variance – where there might be equally valid interpretations that account for differences in the error classification.

This reliability allows us to share the coding effort across several researchers and many source texts. That is, the results are broadly similar across researchers and can, through the automatic translation process that we have adopted, be presented in the appropriate form for the specific purposes of the different software developers. Again, since we have different motivations for the use of the error analyses, we have to admit the possibility that different researchers would consider an error differently when coding the source texts. The outcome of the analyses yields important information about the patterns and relative frequency of errors made by te reo Maori language learners. It allows us to concentrate effort on providing support through computer-based tools on the most frequently occurring errors and to decide how appropriate support could be provided.

Conclusion

We have described a procedure of identifying errors from student assessments and a general-purpose classification scheme for arranging the errors. This scheme is flexible enough to allow data to be recorded for several different purposes, both for teaching and learning support directly related to the assessments and other activities such as building linguistic models and software development.

We are confident that the classification can be used reliably between different individuals to analyse written student work. The results from the pilot study indicate the level of reliability that can be achieved between researchers using the classification and procedure for marking.

Our main practical conclusion is that anyone running a course based on the first three chapters of *Te Kākano* might expect to encounter a pattern of errors similar to those that we found in our study. Some of the errors we have seen are likely to be found in all learners of Maori at this level of development.

We expect that the classification scheme will undergo further refinement as we look at a greater variety of assessments. This work is expected to continue within our own research groups. Furthermore, by making the classification system available more

widely, we hope to enhance its development through the contribution of others working in similar areas.

It seems logical that any kind of exercises that targets the correct usage of macrons, determiners, ā/ō, possessive pronouns and kinship terms should see a reduction in the total number of errors made. Exercises that target the difference between short answer formats and full answer formats would also be beneficial. However, our intention is not to develop a teaching strategy that eliminates errors altogether. After all, errors are a natural, indeed necessary, part of language learning. Our intention in developing this classification is to create a diagnostic tool for the purpose of classifying errors and, through a process of instructional design, to link this system to useful exercises that will support students in the further process of learning.

There is more work required to validate the procedures we use to help to ensure that different assessors can code errors with the same degree of reliability. The error classification in its present form consists of a set of classification codes and examples of error for each class. As such, it represents an outline marking scheme, one that is particularly suited to the *Te Kākano* source text. We would like to encourage other researchers and teachers working in similar areas to consider using this classification system in their work. This would greatly assist in its further development.

For other texts, and to allow the classification scheme to be more useful to a wider body of researchers, the process of assessment can be coordinated amongst a group of assessors. By using the error classification scheme as an outline, and then conducting and discussing a sample marking exercise, the assessment team can evolve their own specific marking schemes to suit their own purposes. By computing kappa statistics for the cross-marked sample, the assessment team will be able to calibrate their marking exercise and identify any potential variance among the team.

In future work, the Computational Linguistics group will use the error analysis to develop a computational grammar and an automated human-machine dialogue system, both of which are sensitive to the most commonly occurring errors. The Software Engineering group will use the error distributions discovered through this analysis as the starting point for the design of instructionally-centred exercises and software systems to support student learning. Further analysis and evaluation of these systems will be carried out to determine their effect on the number and type of errors that students typically make.

References

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics, 2*(249-254).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement, 20*, 37-46.

Corder, S. P. (1967). The Significance of Learners' Errors. *International Review of Applied Linguistics in Language Teaching, 5*(4), 161-172.

Crombie, W., Houia, W., & Reedy, T. (2000). Issues in Testing the Proficiency of Learners of Indigenous Languages: An example relating to young learners of Maori. *Journal of Maori and Pacific Development, 1*(1), 10-26.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: a response to Truscott. *Journal of Second Language Writing, 8*(1), 1-11.

Houia, W. (2002). An Analysis of Typical Errors of Young Learners of te reo Maori. *Journal of Maori and Pacific Development, 3*(1), 44-72.

Johnson, D. and Rolleston, A. (2001). In-service provision for teachers of Maori language and teachers who teach through the medium of Maori: A working model reviewed. *Journal of Maori and Pacific Development, 2*(1), 21-32.

Moorfield, J. C. (2001). *Te Whanake 1: Te Kakano*. Auckland, New Zealand: Addison Wesley Longman.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369.