Tagging L2 Writing: Learner Errors and the Performance of an Automated Part-of-Speech Tagger

Roslina Abdul Aziz
leenaziz@pahang.uitm.edu.my
Universiti Teknologi MARA Cawangan Pahang, Malaysia

Zuraidah Mohd Don
zuraidah.mohddon@utm.my
Universiti Teknologi Malaysia

ABSTRACT

This paper is concerned with the application of technologies developed in other disciplines, in particular with the use of text processing techniques to investigate the problems of second language learner writing in English. The question addressed is whether learner texts produced by L1-Malay learners at the University of Malaya can usefully be processed using the Constituent Likelihood Automatic Word-tagging System (CLAWS); a part-of-speech (POS) tagger developed for and trained on texts written by native speakers of the language. The study adopts the procedure employed by van Rooy and Schäfer (2002). CLAWS was used to automatically POS tag a subset of the Malaysian Corpus of Learner English (MACLE), and the texts were then analyzed for tagging accuracy. CLAWS was found to perform less well on learner text than on native speaker texts, but still with an accuracy rate of over 90%. The sources of error are traced, and spelling errors are found to be the most common source. Closer inspection indicates that successful tagging is likely to lead to problems downstream in later processing, which suggests that to optimize performance, some modifications will be required in tagger design.

Keywords: Learner Corpora; Learner Errors; Part-of-Speech Tagging; Tagging Accuracy

INTRODUCTION

Collaboration across traditional disciplines has in recent decades enabled technologies developed first in computer science and then in computational linguistics to be applied in language education. Corpus linguistics have benefitted tremendously from these computational technologies, in particular the development of various computational tools and softwares specifically developed in aiding the process of developing, storing and analyzing the corpus data. Techniques developed in corpus linguistics complement traditional methods make it possible to investigate collectively the learning problems of large groups of students. Corpus linguistic methodology begins with the compilation of texts or corpus written by a group of learners, and natural language processing techniques are used to analyse the texts to find out what the learners can and cannot do in the target language.

In the study of learner data, some limited information can be obtained using conventional search facilities. For example, to find out how learners use BE, we can make separate searches on be, was, are, being, etc. To make use of this information, however, we really need to know something about the words to the left and to the right. The solution is known as part-of-speech (POS) tagging, the most frequently used form of annotation for learner corpora (Leech, 1997) which indicates the word class of each word, and facilitates searches according to word class and word class combinations.
The storing of texts in electronic format enables the use of available software tools, thus making it possible to “quantify learner language”, “uncover interlanguage patterns of use” and “enrich learner data with a wide range of linguistics annotation” (Granger, 2008, p. 340). This study investigates the extent to which an automatic tagger such as the Constituent Likelihood Automatic Word-tagging System or “CLAWS” (Garside & Smith, 1997) can profitably be used to tag L2 texts in the Malaysian Corpus of Learner English (MACLE). The aim is to facilitate useful searches to answer significant questions about language learner performance. Following the procedure proposed by van Rooy and Schäfer (2002), the study addressed the following research questions:

RQ1: How accurate is CLAWS in assigning POS tags to Malaysian SL learner writing?
RQ2: What is the effect on the performance of CLAWS of learner errors in MACLE?

LEARNER CORPORA, TAGGING AND LEARNER ERRORS

Although useful information can be extracted from learner unannotated corpora, as attested by Aijmer (2002) and Nesselhauf (2009), “a much richer information can be extracted from corpora that have been annotated” (Diaz-Negrillo & Thompson, 2013, p. 13). Learner corpora are historically associated with error annotation combined with computer-aided analysis (Granger, 2002). Many learner corpora including the International Corpus of Learner English or “ICLE” (Granger, 1993, 2003, 2005) and the two-million-word Japanese Learner English corpus (Izumi, Uchimoto & Isahara, 2005) were tagged for learner errors.

In recent years, the focus has shifted from error tagging to part-of-speech (POS) tagging. The POS annotation of learner texts, however, has received limited attention in the literature. Early attention focused on ICLE, tagged using the Tools for Syntactic Corpus Analysis (TOSCA) set of 220 tags. de Haan (2000) evaluated the performance of the TOSCA-ICLE tagger (Aarts, van Halteren & Oostdijk, 1998) on corpora of Czech, Dutch and Spanish learners of English, and achieved an accuracy rate of 95% for the Dutch learners and spelling errors was identified as one of the main problematic areas. Meunier and de Mönnink (2001) evaluated the performance of TOSCA-ICLE on Dutch and French learners of English, and traced tagging problems to mostly spelling errors.

van Rooy and Schäfer (2002) compared the performance of three automatic taggers, namely CLAWS, TOSCA-ICLE, and the Brill tagger (Brill, 1999), on the Tswana Learner English Corpus (TLEC), and found that accuracy was affected by errors in spelling, lexical choice, verb conjugation, clause type, the use of the infinitive and omissions. CLAWS emerged as the best performer, achieving 96% on unedited data, in comparison to 87% and 89% for TOSCA-ICLE and Brill respectively. The CLAWS result equals its performance on the British National Corpus (Garside & Smith, 1997), which was around 96-97% accuracy in coding the 100 million-word British National Corpus, suggesting that despite being trained on native speaker data, it can also be used profitably to tag L2 learner data.

All three studies identified spelling errors as a major source of tagging problems. deHaan (2000) divided spelling errors into word errors and space errors. Word errors are in turn divided into non-words and real-word errors, while space errors are attributed to missing and extra space. vanRooy and Schäfer (2002) reported that real word errors have greater adverse effects than non-word errors, while space errors almost always lead to problems in tagging. The guessing procedures of automatic POS taggers might be able to predict tags for non-words, but they would not function as well with real-word errors and with space errors.
Díaz-Negrillo, Meurers, Valera and Wunsch (2010) investigated the most appropriate form of linguistic annotation for learner corpora, observing that not being trained on learner data, automatic taggers such as CLAWS are not designed to cope with the ill-formed structures found in a learner corpus, and may not perform well as a result. They tested three different taggers, namely the TreeTagger, TnT and the Stanford tagger, on a 39015-word section of the Non-native Spanish Corpus of English-NOCE (Díaz-Negrillo & Garcia-Cumbreras, 2007), which was tagged manually for learner errors. Several major mismatches, all due to learner errors, were reported in the POS classification variables, namely stem, distribution and morphology. These mismatches were valuable since they gave new insights into the grammatical properties of learner errors. Diaz-Negrillo et al. (2010) proposed a threefold split in POS information to improve the insight provided by the tagging, and interlanguage POS annotation is believed to overcome the problem caused by the use of a single POS tagger trained on native speaker data.

The literature identifies learner errors as a problem in learner corpora that needs to be addressed before undertaking tagging or other annotation. Although the evidence so far indicates that learner errors can impair the accuracy of tagging, there is still a need for more studies in view of the small number and the very limited scope of studies carried out so far (Díaz-Negrillo et al., 2010). The focus has been on a very limited set of corpora including ICLE, TLEC and NOCE, which together make up only a small and not necessarily representative sample of learner language. Studies are needed on learner corpora from other parts of the world involving other L1 backgrounds and different experiences of exposure to English, so that wider comparisons can be made of the effects of different kinds of learner errors. This study seeks to provide empirical evidence of the influence of learner errors on the accuracy of automatic tagging using the written data obtained from the L1-Malay English learners from Malaysia. The findings will be compared to those discussed in previous studies, and attention will be paid to the effect of these errors on the accuracy of automatic tagging.

THE GRAMMATICAL TAGGING OF ESL LEARNER TEXTS

The CLAWS tagsets, which are much larger than the conventional set of eight parts of speech, were developed for texts produced by native speakers of the language, and trained on native speaker corpora. Although they are not designed to recognise and annotate learner errors (Díaz-Negrillo et al., 2010), their use is an appropriate first step in the analysis of learner texts. Since taggers are designed to work on grammatically well-formed texts, and the success rates of established taggers are known, grammatical tagging is a logical first step in the processing of texts known to contain learner errors.

Before annotating the selected subset of the MACLE (Knowles & Zuraidah, 2004; Knowles et al., 2006), and before relying on further research on the tags assigned, we have to ascertain how accurate the tagsets tag learner corpora, how their accuracy is affected by learner errors, and to what extent their shortcomings are likely to distort the findings. Consider, for example, the excerpt below taken from MACLE, which contains three sentences:

(1) They just need to make tesis to finish their studies.
(2) So, they are not exposure with real world out there.
(3) Students should give exposure fir the real world in their studies.

There are several learner errors:
i. spellings such as “tesis” (thesis), and “fir” (for);
ii. lexical choices such as make instead of produce, write, or complete;
iii. the misuse of prepositions including with instead of to;
iv. inappropriate grammatical class in the use of exposure in the place of exposed.
v. omissions including articles; a before “tesis”, and the before real world out there.

We tagged the text (Sample B0017-05) using CLAWS4 as illustrated below.

They_PNP just_AV0 need_VVB to_TO0 make_VVI tesis_NN1 to_TO0 finish_VVI their_DPS studies_NN2 ._SENT -----_PUN
So_AV0, PUN they_PNP are_VBB not_XX0 exposure_NN1 with_PRP real_AJ0 world_NN1 out_AVP there_AV0 . _SENT --
----_PUN Students_NN2 should_VM0 give_VVI exposure_NN1 fir_NN1 the_AT0 real_AJ0 world_NN1 in_PRP their_DPS
studies_NN2 . _SENT -----_PUN

CLAWS coped with these errors, generating appropriate tags in most cases despite the errors in the text, and generating two misleading tags from errors of a kind the tagging algorithm was not designed to handle.

If just in they just need were to be tagged as an adjective as in a just cause, there would be something wrong with the procedure that assigns tags, as CLAWS was designed using the probabilistic and rule-based elements. The probabilistic element in the tagger enables it to select a grammatical or part-of-speech (POS) tag for a word by calculating the likelihood of all the probability of all possible tags to occur in a particular context and choose the tag sequence with the highest probability (Garside & Smith, 1997). CLAWS generates no errors at all in this sense, and any bizarre results are caused by the nature of the data. For example, the internal word list accessed by CLAWS will indicate that fir and exposure can only be nouns. Both words were tagged correctly according to their spellings as singular nouns [NN1]. Unlike CLAWS, a human reader has access to syntax and the meaning of the text, and will realise that the expression give exposure cannot be appropriately used in its context, and that the text has nothing to do with fir trees. Although give exposure is correctly tagged in its immediate context as infinitive verb and noun, what is required here is be and a past participle. Fir is clearly a spelling error for for, which is in any case incorrectly used in place of to. Grammatical tagging is a useful first step, but it leaves the deeper analysis of the text incomplete.

The study reported here is part of a larger project which aims to find out from a tagged version of MACLE how BE is used by L1-Malay learners in a Malaysian university. However, the research soon encountered a methodological obstacle, because in view of the prevalence of learner errors it was impossible to know whether the tags assigned by CLAWS as the selected tagger would reach the level of accuracy acceptable or comparable to what it would obtain in tagging the native speaker data. The preliminary study of the tiny sample above yielded encouraging results, but before proceeding, it was necessary to test CLAWS on a larger sample.

**METHODOLOGY**

**CORPUS DATA AND TAGGING**

The data for this study was taken from the Malaysian Corpus of Learner English or MACLE (Knowles & Zuraidah, 2004; Knowles et al., 2006), which is a learner corpus consisting of approximately 800,000 words of argumentative essays written by second to fourth year
students from the University of Malaya. The essays were written by learners from three different L1 backgrounds: Malay, Chinese and Tamil. For this research, only essays written by L1-Malay learners were selected from the L1-Malay sub-corpus consisting of 366 essays of about 500 words, and amounting in total to 198,262 words. A 10% sample of 36 essays was selected from the sub-corpus and tagged using CLAWS. Only essays containing learner errors were selected (refer to Table 1 for the list of errors) because CLAWS is known to tag grammatically well-formed structures with great accuracy, and the aim was to find out if the same accuracy can be obtained with ill-formed structures.

The automatic tagger used was CLAWS, developed by UCREL at Lancaster University (Garside & Smith, 1997). CLAWS was chosen because it is a hybrid tagger combining probabilistic and rule-based procedures and on account of its successful record. The probabilistic approach is necessary because many English words are tagged differently in different contexts, e.g. round is a preposition in round the corner and a noun in a round of golf. CLAWS uses the probabilities of tags and tag sequences to select from all possible tag sequences the tag sequence with the highest probability in the context (Garside & Smith, 1997). The probabilistic element alone, however, is insufficient to tag a text accurately as it “treats the tag sequence as an abstraction” (Garside & Smith, 1997, p. 105). It is therefore unable to assign exceptional coding accurately for expressions such as idioms, multiwords or foreign expressions.

CLAWS incorporates a rule-based component to complement the probabilistic component, and so is able to tag idioms such as as well as as single token and compounds such as dining room as NOUN-NOUN rather than ADJECTIVE-NOUN. The inclusion of both probabilistic and rule-based components has enabled CLAWS to achieve 96% - 97% accuracy in tagging the 100 million-word British National Corpus (BNC) and other texts (Garside & Smith, 1997). CLAWS has a long track record beginning with LOB and COLT (Garside & Smith, 1997), and more recently is reported to be the most accurate (96%) in tagging the unedited Tswana Learner English Corpus (TLEC), and is considered the university benchmark (van Rooy & Schäfer, 2002).

Although CLAWS is not generally available for free use, a trial tagging service for a limited set of 300 words is available for free on the UCREL website. In view of the relatively small size of the sample, it was possible to use this facility by tagging the selected texts in small batches. The tagset used was C7, which contains over 160 tags, and is the enriched version of C5, which only has 60 tags. C7 is considered the current standard and is available online at the UCREL website.

ERROR CATEGORIES

In evaluating the accuracy of CLAWS to tag L2 learner writing, a distinction has to be made between misleading tags assigned as a result of learner errors and those caused by residual inaccuracies in the tagging procedure itself (van Rooy & Schäfer, 2002). This study focuses on tagging errors brought about by learner errors.

The learner errors were initially classified using a system of ten broad categories devised by van Rooy and Schäfer (2002), but the system had to be extended to include errors arising from overgeneration (Ionin, 2011, 2002), word order and word form. Overgeneration errors constitute the inappropriate insertion of BE before a main verb to produce ill-formed constructions such as is come as in knowledge about the job is come from theoretical (refer to Rosлина & Zuraidah, 2014). Word form errors involve the use of words of the wrong grammatical class, for instance the adjective confident instead of the noun confidence; while word order errors involve the inappropriate ordering of a phrase or clause, e.g. student university instead of university students. The decision to include these three additional categories was made after observing the recurrence of errors belonging to these categories in
the data. It is also important to note that in contrast to van Rooy and Schäfer (2002), this study includes all instances of omissions including the omission of lexical items, verbs, prepositions and articles in the same category of omission errors. Table 1 below summarises the tag error categories used in this study.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Lexical choice-wrong lexical item</td>
<td>*They revolve in club and society…</td>
</tr>
<tr>
<td>2 Verb conjugation-wrong tense, aspect, and/or number</td>
<td>*…there is various subjects offered by…</td>
</tr>
<tr>
<td>3 Article-wrong or superfluous article</td>
<td>*…in a real world practical knowledge…</td>
</tr>
<tr>
<td>4 Number in noun phrase-incongruency between singular and plural</td>
<td>*…many of the graduate student…</td>
</tr>
<tr>
<td>5 Clause type-mainly independent, finite clause where dependent/non-finite clauses should be used</td>
<td>*…poverty is the cause people in africa are very poor…</td>
</tr>
<tr>
<td>6 Resumptive pronoun</td>
<td>*…another factor is about the silibus they the university prepared…</td>
</tr>
<tr>
<td>7 Infinitive-inappropriate use, or non-use where appropriate</td>
<td>*…what they want is just finish up their…</td>
</tr>
<tr>
<td>8 Omission of word</td>
<td>*…not ___ of the courses in university nowadays…</td>
</tr>
<tr>
<td>9 Preposition-wrong preposition</td>
<td>*…expose them with the real world…</td>
</tr>
<tr>
<td>10 Spelling-non-word, real-word, missing space, extra space and words borrowed from L1</td>
<td>*…people that are low standard than them…</td>
</tr>
<tr>
<td>11 Overgeneration- inappropriate insertion of BE before a main verb</td>
<td>*…when they are graduate…</td>
</tr>
<tr>
<td>12 Word form-wrong form of the intended word</td>
<td>*In the same time the level of confident …</td>
</tr>
<tr>
<td>13 Word order- incorrect arrangement of words in a phrase, clause or sentence</td>
<td>*Student university often have to…</td>
</tr>
</tbody>
</table>

TABLE 1. Error categories

Adapted from van Rooy and Schäfer (2002, p. 331)

ANALYSIS PROCEDURE

The study adopts the procedure employed by van Rooy and Schäfer (2002), with some modification to the size of the sample data and the number of automatic taggers used to tag the data. van Rooy and Schäfer (2002) used CLAWS7, TOSCA-ICLE and Brill to analyse five sample texts, but we decided to tag a larger sample consisting of 36 scripts, amounting to about 10% of the L1-Malay learner sub-corpus, using a single tagger, namely CLAWS4 with the C7 tagset. When tagging was completed, the tagged files were edited manually to identify and correct tagging errors, and the edited versions were stored in a separate file.

It is important to note that whereas van Rooy and Schäfer (2002) corrected spelling errors and re-tagged the words, in this study errors were left as they were, so that spelling errors and words borrowed from the learners’ L1 were left. However, this study follows van Rooy and Schäfer (2002) in distinguishing tagging errors in just two categories, namely those caused by learner errors and those made by the tagger itself, and ignores the shared-blame category (learner error + tagger error) proposed by de Haan (2000) and Meunier and de Mönnink (2001). The reason for this is that van Rooy and Schäfer (2002) found it difficult to distinguish errors resulting solely from learner errors from shared blame errors.
RESULTS

This section reports and discusses the results of the analysis of potential learner errors. It begins by presenting the data from all error categories, followed by a detailed presentation of spelling errors, and ends with the overall tagger accuracy.

LEARNER ERRORS

Table 2 presents the distribution and frequency of the approximately 1986 cases in the tagged texts which were judged to exhibit learner errors. Errors in spelling, number concord, verb conjugation, omissions and the misuse of prepositions were found to be the most frequent categories. Spelling errors were most frequent at 22%, followed by verb conjugation and preposition misuse at about 17%, and omission errors at about 15%.

<table>
<thead>
<tr>
<th>No</th>
<th>Category</th>
<th>All Errors</th>
<th>Tagging Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tokens</td>
<td>%</td>
<td>Tokens</td>
</tr>
<tr>
<td>1.</td>
<td>Lexical choice</td>
<td>99</td>
<td>5.0</td>
</tr>
<tr>
<td>2.</td>
<td>Verb conjugation</td>
<td>346</td>
<td>17.4</td>
</tr>
<tr>
<td>3.</td>
<td>Article</td>
<td>78</td>
<td>4.0</td>
</tr>
<tr>
<td>4.</td>
<td>Number in noun phrase</td>
<td>345</td>
<td>17.4</td>
</tr>
<tr>
<td>5.</td>
<td>Clause type</td>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>6.</td>
<td>Resumptive pronoun</td>
<td>20</td>
<td>1.0</td>
</tr>
<tr>
<td>7.</td>
<td>Infinitive</td>
<td>40</td>
<td>2.0</td>
</tr>
<tr>
<td>8.</td>
<td>Omission</td>
<td>294</td>
<td>14.8</td>
</tr>
<tr>
<td>9.</td>
<td>Preposition</td>
<td>115</td>
<td>5.8</td>
</tr>
<tr>
<td>10.</td>
<td>Spelling</td>
<td>436</td>
<td>22.0</td>
</tr>
<tr>
<td>11.</td>
<td>Overgeneration</td>
<td>71</td>
<td>3.6</td>
</tr>
<tr>
<td>12.</td>
<td>Word form</td>
<td>96</td>
<td>4.8</td>
</tr>
<tr>
<td>13.</td>
<td>Word order</td>
<td>30</td>
<td>1.5</td>
</tr>
<tr>
<td>14.</td>
<td>Others</td>
<td>6</td>
<td>0.3</td>
</tr>
<tr>
<td>Total</td>
<td>1986</td>
<td>100</td>
<td>1117</td>
</tr>
</tbody>
</table>

As seen in Table 2, errors in articles, prepositions, clause type, resumptive pronouns, infinitives, overgeneration, word order and number in noun phrases were found to have almost no effect on tagger accuracy. According to van Rooy and Schäfer (2002), articles and prepositions require very little disambiguation and they belong to a very simple syntactic category, so that if an article or preposition is used inappropriately, it is still tagged correctly. Preposition errors, although frequent, contribute only about 0.04% of tagging errors. Errors in articles do not affect tagger accuracy at all. Extract (1) shows samples of errors in preposition of and article the, which were appropriately tagged as a preposition [IO] and an article [AT] respectively.

(1) …Many_DA2 of_IO people_NN are_VBR agree_VV0 that_CST the_AT degree_NN1 will_VM make_VVI their_APPGE life_NN1 comfortability_NN1 in_II their_APPGE future_NN1._. B0068

Nouns and pronouns are tagged for singular and plural. As long as they are used in the correct syntactic position they are tagged correctly according to the forms used (van Rooy & Schäfer, 2002). As exemplified in Extract (2) below, the use of singular seminar (in bold) was tagged as a singular, even though the preceding determiner many clearly indicates that
the noun should be plural. The nouns underlined are also judged to be plural, but it is difficult to be absolutely certain as they could have intended to be singular, in which case the verbs involve and study were inappropriately used. In view of the ambiguity of these nouns, the tags were not regarded as errors.

(2) Student_NN1 that_CST involve_VV0 in_II these_DD2 activities_NN2 had_VHD most_DAT advantages_NN2 and_CC opportunity_NN1 than_CSN the_AT student_NN1 that_CST just_JJ study_NN1 for_IF the_AT examination_NN1 __. There_EX are_VBR many_DA2 seminar_NN1 and_CC workshops_NN2 that_CST conducted_VVD by_II university_NN1 for_IF their_APPGE student_NN1 __B0038-05

Nearly 17% of the learner errors involve number agreement in the noun phrase, but they produce only about 0.04% of tagging errors, which confirms that errors in number have a minimal effect on tagger accuracy. Omissions and errors in spelling and verb conjugation seem to affect the performance of CLAWS more substantially. As in previous studies (de Haan, 2000; Meunier & de Mönnink, 2001; van Rooy & Schäfer, 2002), spelling errors had an adverse effect on accuracy, and were associated with 1.95% of inaccuracies. Two types of spelling error occur most frequently, namely those that produce a non-existent word, and those that produce a real English word but the wrong one, for example two is spelt as too (deHaan, 2000; van Rooy & Schäfer, 2002). Non-word spelling errors are generally not a problem since automatic taggers have guessing modules to assign tags to items not in the lexicon (van Rooy & Schäfer, 2002). Real-word errors, however, can cause more serious problems when the intended word belongs to a different grammatical class, as in the case of doe, which is a noun in contrast to the auxiliary do. This applies to the majority of the real word errors found in the MACLE texts. Spelling errors are discussed further in the following sub-section.

Errors of omission are found to contribute 1.82% of all errors. The reason for this is straightforward. All missing words in the data were not tagged simply because CLAWS was not programmed to detect and tag missing words. Interestingly, omission led to tagging errors for the preceding word. As shown in Extract (3) the symbol “∅” indicates a missing article the, which has resulted in right being tagged as an adverb [RR] instead of a noun [NN1]. It is important to explain here that omission errors constitute all cases of covert items from all parts of speech, for instance the omission of the is categorised below as an omission error, and not as an error in the use of the article.

(3) In_II south_ND1 Asia_NP1 like_II Bangladesh_NP1 __, India_NP1 and_CC west_ND1 country_NN1 __, women_NN2 have_VH0 ∅ right_RR to_TO fight_VVI for_IF similarity_NN1 __PJ0002-05

Verb conjugation errors also contribute substantially to tagging errors (1.54%). Most of the cases involve the use of unmarked verbs in place of marked verbs as exemplified by the verb give in Extract (4). In this case the tagger assigns a tag for the unmarked verb (VV0 – base form of lexical verb), when it should have been tagged as marked [VVZ]. According to van Rooy and Schäfer (2002) a probabilistic aspect of an automatic tagger that was trained on native data would not be able to tag such a form accurately, as it would not have occurred during training.

(4) In_II same_DA time_NNT1 __, the_AT pollution_NN1 give_VV0 the_AT affect_NN1 the_AT activities_NN2 in_II the_AT town_NN1… __H0004
This study also included another type of learner error, namely overgeneration errors (Ionin & Wexler, 2001, 2002; Roslina & Zuraidah, 2014). Although, overgeneration errors account for only (3.4%) of errors, when they do occur they inevitably lead to tagging errors. As shown in Extract (5), the insertion of *is* has affected the tag assigned to the preceding main verb *know*, which should be tagged [VVN] since the main verb is positioned after a *be* as in *be + PP*, therefore, should be tagged as [is_VBZ know_VVN], as *be + bare V* [VBZ + VVO] is never allowed in English syntax.

(5) Students_NN2 who_PNQS is_VBZ know_VV0 everything_PN1 , but_CCB if_CS we_PPIS2 give_VV0 him_PPHO1 a_AT1 project_NN1 to_TO handle_VVI it_PPH1 ,_, K0017

One important observation made from the ill-formed constructions such as overgeneration of *be* and omission, is that they almost always lead to tagging errors. These non-canonical constructions are not limited to *be* overgeneration and omission errors, for learners are also observed to use the ill-formed combination *modal + inflected V*, in which case the lexical verb is inflected with *–ing/*-ed/-es*. This confuses the tagger and most probably results in a tagging error. Extract (6) below is an example *can* being tagged [VVO] or as a base form of a lexical verb instead the correct tag [VM] for verb modal. In addition, the main verb should be tagged as an infinitive, since *modal + Ving* is not syntactically possible in English.

(6) People_NN before_I us_PP202 ,_, maybe_RR about_RG 100_MC or_CC several_DA2 hundred_NNO years_NNT2 ago_RA ,_, they_PPHS2 can_VV0 dreaming_VVG because_CS they_PPHS2 live_VV0 in_II peaceful_JJ place_NN1 ._.. K0030

The same effect is also observed with non-target constructions involving words with two or more grammatical functions such as *to*, which can mark the infinitive or function as a preposition. In Extract (7) below, *to* is inappropriately used in place of *of* (instead of), while the writer has also inappropriately used the word *get* in the place of *getting*, since *instead of* is always followed by a noun or gerund. Because *to* is followed by uninflected *get*, the two words were treated as an infinitive, *to* being assigned the tag [TO] and the verb the tag [VVI]. Note also *instead* was tagged as an adverb [RR] rather than a preposition [II] since it was not followed by *of*.

(7) As_CSA conclusion_NN1 ,_, instead_RR to_TO get_VVI a_AT1 knowledge_NN1 and_CC education_NN1 the_AT system_NN1 must_VMchange_VVI ,_, lecturer_NN1 must_VMgive_VVI what_DDQ they_PPHS2 have_VH0 to_II the_AT student_NN1._. B0041-05

Extracts 3, 5, 6 and 7 show how ill-formed constructions influence tagger performance, and most importantly, the tagger is unable to recognise these structures and tag them appropriately.

**SPELLING ERRORS**

Misspelling, according to de Haan (2000) and van Rooy and Schäfer (2002), is one of the major sources of inaccuracy in tagging. As shown in Table 3, a total of 436 spelling errors were recorded in 16138 word tokens, a frequency of approximately 27 spelling errors per 1000 words.
TABLE 3. Categories for spelling errors

<table>
<thead>
<tr>
<th>Category</th>
<th>All Errors Tokens</th>
<th>%</th>
<th>Tagging Errors Tokens</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Non-word – the result of misspelling of a word that doesn’t exist e.g. the country’s requiremat is not…</td>
<td>144</td>
<td>7</td>
<td>75</td>
<td>52</td>
</tr>
<tr>
<td>2 Real-word - the result of the misspelling is a different real word of English e.g. there are may business in this…</td>
<td>183</td>
<td>8.9</td>
<td>171</td>
<td>93</td>
</tr>
<tr>
<td>3 Borrowed words - the result of borrowing words from learners’ L1/ using the L1 spelling of a word e.g. to give a money emaskahwin (dowry)…</td>
<td>14</td>
<td>0.7</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>4 Missing space- two words written as one e.g. to check your bankaccount number…</td>
<td>17</td>
<td>0.8</td>
<td>15</td>
<td>88</td>
</tr>
<tr>
<td>5 Extra space- a single word written as two words e.g. damaging the rain forest</td>
<td>78</td>
<td>3.8</td>
<td>46</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>436</td>
<td>21.2</td>
<td>314</td>
<td>72</td>
</tr>
</tbody>
</table>

deHaan (2000) divided spelling errors into two broad categories; word errors and space errors. Word errors are divided into non-words and real-words, and the category also includes capitalisation errors. Non-word errors result in a word that does not exist such as “teorycal” or “studats”, while real-word errors result in a different real word (de Haan, 2000; van Rooy & Schäfer, 2002); for example “loss” for lost, “doe” for do, or “fir” in place of for. Capitalisation errors were almost non-existent, and were consequently excluded. There were quite a number of borrowed words, and these were classed as word errors.

Space errors involve missing spaces and extra spaces. Missing space errors result in two or more words written solid as one, for example “bankaccount” for bank account and “workhard” for work hard. Extra space errors divide single orthographic words into two, for example “can not” for cannot, and “with out” for without.

As shown in Table 3, almost 17% of learner errors constitute word errors, almost 16% were contributed by non-word and real-word errors. Non-word errors are generally not a problem, since automatic taggers are able to use guessing procedures to assign tags to words that are not in their lexicons, provided that the non-word is placed in the correct syntactic position (van Rooy & Schäfer, 2002). For example, the word “graduatas” (for graduates) in Extract (8) below is tagged [NN2] (plural noun) despite the spelling error.

(8) The _AT_ unemployed _JJ_ graduatas _NN2_ are _VBR_ those _DD2_ who _PNQS_ do _VD0_ not _XX_ realize _VVI_ that _DD1_ academic _JJ_ qualifications _NN2_ are _VBR_ not _XX_ sufficient _JJto_TO obtain _VVI_ a _AT1_ job _NN1_: H0010

These guessing procedures, however, can be adversely affected by omission errors, as can be seen in Extract (9), from which a noun such as students or graduates is missing, its absence being indicated by “∅”. The omission has caused two thirds of the non-words to be inaccurately tagged. The noun “medice” (presumably medical) was accurately tagged [NN1]-singular noun, while “macancal” (presumably mechanical), was assigned the inappropriate tag [JJJ]-adjective, and “engineering” (engineering) was tagged [VVG]-ing as the participle of a lexical verb, instead of [NN1].
But the government agencies require personally qualified economics, mechanical, medical and engineering (students/graduates).

There are, however, instances of non-words inaccurately tagged despite being appropriately positioned syntactically, as exemplified by the word “amang” for among in Extract (10) below. The intended preposition, instead of being tagged [II]-general preposition, was tagged as a base verb form [VV0]. This indicates that the total dependency on the guessing module to tag non-word errors needs to be reconsidered by researchers working with L2 data. Approximately 52% of the non-word errors in this study contribute to tagger errors, indicating that despite its stability as attested by van Rooy and Schäfer (2002), the accuracy of CLAWS for tagging non-word errors would be seriously affected by other learner errors as shown in Extract (9). More importantly, the tagger would also tend to tag a non-word incorrectly even when it is positioned in an ideal location, as exemplified in Extract (10) below.

(10) So RReq the unemployment problem is mainly amang the arts graduates.

Unlike non-word errors, real-word errors almost always lead to tagging errors. As found by van Rooy and Schäfer (2002) and de Haan (2002), misspellings of this type strongly affect tagging accuracy, as can be seen from Table 3, which shows that about 93% of real-word errors lead to tagging errors. In most cases, errors occurred when the observed and intended words belong to different grammatical classes, as exemplified in Extract (11) below. The misspelled “may” for many is a modal verb and so tagged [VM]-modal verb instead of [DA2]-plural after-determiner. The same applies to “loss”, which is a noun, unlike the intended verb lose.

(11) Nowadays, there may business in this country. Because 21 of the money, they attention to their children.

Borrowed word errors are the least likely to occur of all word errors, only 17 incidences occurring in the data, or just 0.7%. Nevertheless, half of these (50%) lead to tagging errors. Being trained on native speaker data, CLAWS was not designed to guess the tags for expressions which include borrowed words (Diaz-Negrillo et al., 2010). In some cases, borrowing extends into code switching, giving rise to expressions which would require the tagger to recognise foreign syntactic rules. Extract (12) below provides an example of tagging errors involving the expression kongsigelap (secret society or underground organisation). The noun phrase includes the noun korsi which should be tagged [NN1], and the adjective gelap, which in a Malay text should be tagged [JJ]. The problem is of course that CLAWS was not designed to deal with postnominal adjectives, and guesses that gelap must be a verb on the grounds that this is the content word class most likely to follow a noun.

(12) The noun phrase includes the noun korsi which should be tagged [NN1], and the adjective gelap, which in a Malay text should be tagged [JJ]. The most appropriate tagging for the borrowed words would be [korsi NN1 gelap JJ], using the same tags as for its English equivalent secret society [secret JJ society NN1]. The problem is of course that CLAWS was not designed to deal with postnominal adjectives, and guesses that gelap must be a verb on the grounds that this is the content word class most likely to follow a noun.
(12) …hired JJ killer NN1 synonymed VVD as II the AT gangster NN1 or CC the AT members NN2 of IO a AT1 criminal JJ gang NN1 or CC well-known JJ as CS A kongsi NN2 gelap VVO in II Malaysia NP1 __ A0017

It would, however, be wrong to assume that the tagger inevitably tags borrowed words incorrectly. Extract (13) provides an example of the borrowed word *fenomena* (a Malay spelling of *phenomenon*) correctly tagged as a noun. The positioning of the word; after an article and before the copula *is*, provides a very clear indication that it is a noun, leaving the tagger with very small margin for an incorrect guess.

(13) The AT *fenomena NN1* is VBV very RG bad JJ or CC difficult JJ for IF our APPGE country NN1 __ F0071

Spelling errors in the second category, namely space errors, are also very likely to cause tagging errors. vanRooy and Schäfer (2002) found that space errors almost always contributed to tagging errors for the three taggers they tested. Even though it was found that both missing space and extra space errors adversely affected tagger accuracy, the former was more likely (88%) than the latter (59%) to cause tagging errors.

As mentioned earlier, missing space errors are very likely to contribute to tagging errors (van Rooy & Schäfer 2002), since compounding two words, of different grammatical classes could result in an English real word or non-word, which could influence tagging accuracy. Take for instance the missing space error “workhard”, which should be written *work hard* and tagged [work VVI hard RR] instead of [workhard NN1] as in Extract (14) below.

(14) …life NN1 is VBV short JJ and CC this DD1 will VM make VVI the AT people NN lazy JJ to TO work VVI and CC do VDI nt XX want VVI to TO workhard NN1 to TO be VBI a AT1 good JJ people NN __ B0060

Extra space errors are found to mostly involve compound nouns mistakenly written as two words as in the case of the adverb *nowadays* in Extract (15) below. As a result of the split, *now* and *days* were tagged as two individual words, instead of just one.

(15) The AT value NN1 of IO family NN1 institution NN1 is VBV also RR decreased VVN now RT days NNT2 __ S0034-05

Another interesting observation is that more than half of the space errors found in this study involve splitting the modal *cannot* and writing it “can not”, as exemplified in Extract (16) below. In this case the tag assigned to *can* [VM] is considered correct, as *cannot* would also be assigned the same tag by CLAWS, since the tagger was not designed to distinguish positive *can* from negative *cannot*. The tag assigned to *not*, however, was considered a tagging error.

(16) The AT market NN1 of IO job NN1 is VBV very RG extensive JJ and CC can VM not XX have VHI problem NN1 not XX have VH0 job NN1 in II Malaysia NP1 __ H0012
THE ACCURACY OF CLAWS

Of the 16138 word tokens in the data, 1986 (12%) were potential learner errors. Further analysis shows that 948 or approximately 6.5% of these errors were incorrectly tagged by CLAWS. The figures suggest that CLAWS reached approximately 93.6% accuracy, which is about 3% lower than the accuracy of between 96% and 97% achieved with the BNC (Garside & Smith, 1997). Table 4 below summarises the overall accuracy of CLAWS.

| TABLE 4. Overall accuracy of POS tagging in MACLE |
|---|---|
| Learner errors | 1986 |
| Tagging errors | 948 |
| Tagger accuracy | 93.6% |

DISCUSSION

This discussion section first considers the findings in the immediate context of the research paradigm, and then takes a step back and considers the findings in a wider context.

ACCURACY IN TAGGING A LEARNER CORPUS

Learner errors, as shown by the results of this study and consistent with the findings of previous studies (de Haan, 2000; Meunier & de Mönnink, 2001; van Rooy & Schäfer, 2002), can significantly affect tagging accuracy. The study confirms that omission, verb conjugation and spelling errors reduced the tagging accuracy of CLAWS to approximately 93.6%, which is lower than the 96%-97% it achieved with the BNC (Garside & Smith, 1997) and 96% with the unedited TLEC (van Rooy & Schäfer, 2002).

The findings suggest that the use of an automatic POS tagger on a learner corpus, especially one that contains a substantial number of learner errors, needs to be complemented by manual tagging procedure. The problem is that manual tagging requires a considerable amount of time and effort. One way to improve accuracy as suggested by van Rooy and Schäfer (2002) is to correct spelling errors before running the tagger. While this can be done for a small corpus, it would be impractical for the 800,000 words of a large corpus like MACLE. However, even if all the spelling errors were corrected, the tagging accuracy would still be affected by other types of learner error such as omission and verb conjugation errors.

Diaz-Negrillo et al. (2010) proposed an interlanguage annotation involving a threefold split in POS information to improve tagging accuracy, annotated separately for the lexical stem, the distribution, and the morphology of the tokens. The results from such observations enable conflicting evidence to emerge and these conflicts or mismatches provide access to the properties of learner language (Diaz-Negrillo et al., 2010). Considering the effects of learner errors on POS tagger accuracy, the tripartite POS analysis suggested by Diaz-Negrillo et al. (2010) seems to be a possible solution for the tagging of learner corpora. The analysis proposed would be able to improve tagger performance, and more importantly, it involves minimal manual annotation.

The findings of previous and current studies also call for the development of an automatic POS tagger trained to cope not only with native speaker data, but also with learner data. It would have to be trained on a sufficient amount of ill-formed structures, preferably those such as spelling errors which are independent of the learner’s L1. Efforts must be taken to profile these structures, so that they can be used in the development of a comprehensive taxonomy of learner errors, which would be useful in the POS tagger training and the
development of error tagging software (Díaz-Negrillo & García-Cumbreras, 2007). Such a device would, however, have to go beyond what we understand as grammatical tagging, and include stemming and parsing, i.e. analysing words and sentences into their parts.

THE VALUE OF TAGGING A LEARNER CORPUS

Tagging a corpus is not normally an end in itself. We process a learner corpus to find out what learners can and cannot do in the target language, and tagging is the first stage of processing. On the negative side, we need to identify the errors that learners make. Of the tagging problems that have arisen in the tagging of the MACLE data, by far the easiest problem to solve is the problem of spelling errors that generate non-existent words. All these are necessarily unknown to the lexicon used by the tagger. A list of unknown words will include all the errors of this kind, and the spellings can be corrected, and a record kept of the original errors. This simple step alone, will improve the performance of the tagger.

Tagging is followed in natural language processing by parsing, which involves grouping words into grammatical constructions of increasing complexity. The learner data poses difficulties for a non-human parser. A crude parser that just examines the POS tags will build up constructions and ignore the grammatical mistakes that have not affected the word class, resulting in inaccurate tagging. Researchers will in this case still have to scan the texts manually to find the errors. On the other hand, a more refined parser that tests for grammatical well-formedness will probably fail to build up the constructions at all. What is really required for learner texts is a clever parser that builds up constructions while at the same time recognising and recording grammatical errors.

Ill-formed expressions such as *is know* or *they can dreaming* are certainly difficult, but not necessarily impossible to handle. This will require attention to a long-standing methodological anomaly in tagging procedures. Some words are tagged according to their context; for example, the word *transport* is a noun when it follows *the* and a verb when it follows *they*. To examine strings of words is to have syntactic knowledge of a language, and so we have to analyze the syntax in order to identify the tag. The tags are then assumed to be fixed, and they are used to do the syntax. A human reader will correct *is know* to *is known* (which is still wrong) and *can dreaming* to *can dream* before assigning the tags. What is really required is a different procedure in which grammatical class and syntactic structure are analyzed and tagged together. The most difficult cases are those that combine different kinds of learner error, e.g. *may business*, where misspelt *many* is followed by a singular noun in place of a plural one.

The learner problems discussed in this paper all have to do with linguistic form, namely word omission, verb conjugation and spelling errors. As is immediately apparent from the excerpts presented in the Result section, the student writers also had difficulty with the meanings of words and expressions and in using them in appropriate ways, and in finding what the French writer Gustave Flaubert called *le mot juste*. Tagging and parsing will not help in this respect, however, the efficient handling of problems with linguistic form will leave more time to attend to problems of meaning.

CONCLUSION

This paper is concerned with the identification of language learner errors using text processing techniques. It is clear from this research that techniques developed within corpus linguistics can usefully be applied to the study of the formal linguistic problems of language learners. CLAWS as it stands can provide much useful information, and realistic developments in text processing promise much more. Nevertheless, processing facility as
such can only be materialized with the accessibility of if not complete but comprehensive learner error taxonomies. According to Diaz-Negrillo and Fernández-Domínguez (2006) developing error taxonomies is a tremendously complicated process since the taxonomies would “account for diverse dimensions of error classification, encoding conventions and annotation models, which only shows that there is no standard of error” (p. 97). Perhaps at this stage, the interlanguage annotation proposed by Diaz-Negrillo et al. (2010) would suffice.

REFERENCES


deHaan, P. (2000). Tagging non-native english with the TOSCA-ICLE tagger”. In C. Mair& M. Hundt (Eds.), Corpus linguistics and linguistic theory (pp. 69-79). Amsterdam: Rodopi.


Granger, S. (2002). A bird’s eye view of learner corpus research”. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (pp. 3-33). Armsterdam: John Benjamins.


and foreign language education (pp. 337-351). Springer. doi: 10.1007/978-0-387-30424-3_109


ABOUT THE AUTHOR

Roslina Abdul Aziz is currently attached to Akademi Pengajian Bahasa (APB), Universiti Teknologi MARA Cawangan. She received her Ph.D in Corpus Linguistics from University of Malaya, Kuala Lumpur. Her research interests include areas in Corpus Linguistics and Language for Specific Purposes.

Zuraidah Mohd Don is currently an Adjunct Professor at UTM and UPSI, and Chair of the English Language Standards and Quality Council, formed by and under the aegis of the Ministry of Education.