Intelligent Sentence Writing Tutor: A System Development Cycle

Marina Dodigovic, Xi'an-Jiaotong Liverpool University

Abstract. This article focuses on the use of natural language processing (NLP) to facilitate second language learning within the context of academic English. It describes a full cycle of educational software development, from needs analysis to software testing. Two studies are included: 1) the needs analysis conducted to develop the Intelligent Sentence Writing Tutor (ISWT) which diagnoses and at request corrects second language errors in writing, and 2) the summative evaluation of ISWT. The former comprises a survey of learning styles and learner corpus analysis which have both fed into the development of ISWT. The latter is addressed in a quantitative study with elements of both within-sample and comparison design. Due to the interdisciplinary nature of the examined phenomenon, the paper establishes links to previous and concurrent research in the fields of second language acquisition and ICALL (intelligent computer assisted language learning), while relying on a variety of theories and approaches to address a specific educational problem. The comprehensive coverage of the development process takes precedence over the fine detail of individual development stages.

Keywords. Natural language processing (NLP), Intelligent computer Assisted Language Learning (ICALL), Second Language Acquisition (SLA), interlanguage, language errors

THE RATIONALE FOR SYSTEM DEVELOPMENT

The need for an online aid to provide automated feedback on second language (L2) sentence writing arose from despair. An academic English unit which housed a sheltered writing program for non-native speakers of English in Australia faced a challenge of assisting approximately five hundred university students through three courses, numerous workshops and a writing centre facility, all manned with an equivalent of only three full time instructors and no teaching assistants. The demands on the instructors were paramount, the ability to help under the circumstances limited, especially under budgetary constraints, which prevented the hiring of additional teachers. As anticipated through encounters with the literature (Myers, 1997), the students were really interested in feedback on their vocabulary use and grammar. With each student’s one-on-one access to the writing centre capped at one hour per semester, it was hardly possible to deal with individual grammar and vocabulary issues.

Thus necessity gave birth to the idea that students would benefit from an intelligent online tutor, which would provide assistance in self-study mode. Intelligent in this context means emulating the actions of an intelligent human being, so that an impartial observer could get the impression that he or she is communicating with one (Borchardt & Page, 1994). Intelligent language tutors rely on three different computational models: that of language, teacher and student (Heift & Schulze, 2007;
While language is modeled by using grammar and lexis embedded in a parser, i.e. a computer program that analyses student produced language, the teacher and student are most commonly modeled using teaching and learning strategies respectively (Dodigovic, 2005; Heift & Schulze, 2007). Consequently, the development of such a tutor seemed like an ambitious task, involving considerable needs analysis.

The Intelligent Sentence Writing Tutor (ISWT) described here is anchored in the writing curriculum. It affords automated sentence level feedback by attending to the writers’ sentence-level grammar. The main idea underpinning the design of what is essentially a dialogue between ISWT and the student is to ask leading questions on the topic of malaria in order to provide scaffolding for an essay on malaria. The dialogic nature of the interaction makes the learning environment more natural and language use more contextualized (Dickinson, Eom, Lee, & Sachs, 2008). The dialogue is Socratic in the sense that the questions are used to teach content and develop critical thinking skills (Paul, 1995). In order to answer the questions, which start with the more general ones about the disease and then gradually progress to the more specific ones, the student has to be familiar with several readings. The student also has to understand the readings and be able to synthesise the information from multiple sources in his or her own words. Put together, the answer sentences may serve as the basis for an essay.

The above approach corresponds to what Baily and Meurers (2008) called the “Middle Ground” (p. 108) of the language exercise spectrum. Within this range, the student answers are to a certain level predictable, as is often the case with short answers. They are, as the term suggests, in the middle, between the tightly restricted answers of a gap filling exercise and the very loosely restricted essay. This Middle Ground position places the ISWT framework within the “Viable Processing Ground” (Bailey & Meurers, 2008, p.108), where the activity design provides the constraints required for effective language processing and disambiguation (Amaral & Meurers, 2011).

Apart from the content-related restrictions, additional constraints are built into the questions, which through their very structure set the student up for committing one or more of the most frequent language errors identified in student writing. Based on the error type, the ISWT software provides corrective feedback. Whenever the need for language correction arises, the system switches from the topic-related dialog to a remediating tutorial dialog, and returns to the topic when the language issue has been resolved (Weerasinghe, Mitrovic, & Martin, 2009).

The program prototype was developed in Pie-PROLOG and was then adapted for web-based delivery (using Sicstus Prolog and Perl). The development was preceded by extensive needs analysis (Brown, 1995) and followed by evaluation, as mandated by best practice (Decoo & Colpaert, 1999). The former comprises a study of learner language and cognitive styles conducted to develop ISWT, the latter an effectiveness test of ISWT. Both are described in this article in a compact fashion, due to the sheer volume of the process. The entire process spanned several years, being controlled by the availability of funding, equipment and human resources. System development has been informed by theory across disciplines, including second language acquisition, computer assisted language learning (CALL), artificial intelligence (AI), natural language processing (NLP), language teaching, education and software engineering. Consequently, this paper combines and re-contextualises a number of existing theories, models and techniques, resulting in a uniquely new and useful application.
THEORETICAL FRAMEWORK

Learner language

L2 has been recognized as a variable introducing additional complexity into academic writing (Ho, 2007). One of the reasons is that L2 writers tend to have problems with L2 grammar, in particular parts of speech (Yong, 2001) and verb status (Yip, 1995). Such peculiarities of learner grammar are often systemic, and can be referred to as interlanguage (Selinker, 1972; 1997) or learner language (Ellis & Barkhuizen, 2005). The learner languages selected here are the most relevant ones in respect of the study subjects and the learner population for whom ISWT was developed: Chinese-English, Indonesian-English and Arabic-English. This nomenclature implies that the first language (L1) transfer largely shapes interlanguage, a view originally held by Selinker (1972) and initially shared by this author. The strong claim in the course of this project eventually became weakened by the fact that a number of the errors originally identified in one group were later found to be shared by English learners of different linguistic backgrounds. However, this does not diminish the helpfulness of separate interlanguage inventories in shaping ISWT. In fact, it has been suggested that a pedagogical parser used in the error correction process should be trained on learner interlanguage (James, 1998).

Learner language is often replete with the so-called fossilised errors (James, 1998; Selinker, 1972; Pica, 1994). Such errors emerge at various stages of second language development and seem to be very difficult or impossible to eradicate, despite language instruction (Pica, 1994), but appear to be systemic and can therefore be easily classified (Yip, 1995). In Chinese-English Interlanguage, Chang (1987) has identified part of speech confusion in addition to verb form errors, including time, tense and aspect, all due to Chinese not being an inflected language. Yip (1995) finds that Chinese learners mainly have problems with verb transitivity as they use pseudo-passives (*These sentences can analyse many ways*), ergative construction (*What is happened with these verbs?*), tough movement (*Never easy to be learned...*) and existential construction (*There are sentences cause learnability problems*).

Yong (2001) reports that Indonesian based interlanguage has the following features: malformed expressions of feelings/reactions/states based on part of speech confusion (*Parents must take responsible*), missing copula (*Sometimes very easy to make mistake*), finite/nonfinite verb confusion (*I decided to cancelled*).

Another interlanguage that shows evidence of verb related errors is Arabic-English (AbiSamra, 2003; Smith, 2001; Thompson-Panos & Thomas-Ruzic, 1983). The capacity of Arabic to build verbal nouns (Smith, 2001) must add further confusion. Thus, the verb error seamlessly transitions into a part of speech error. Since Arabic groups nouns, adjectives and adverbs into one large part of speech class (AbiSamra, 2003), mutually confusing these parts of speech in English would be expected from a native speaker of Arabic.

Literature thus suggests that the target learner population in this study, consisting mainly of native speakers of Chinese, Indonesian and Arabic, would be expected to have difficulties differentiating between parts of speech in English, with the majority of fossilised errors being linked to English verbs. This insight was used to identify and classify learner errors made by a sample of this learner population. The error classification, based on a corpus of L2 writing, subsequently fed into the content of ISWT.
Error correction
While some members of the L2 writing instruction community believe that errors should be corrected sparingly and according to error gravity (Freiermuth, 1997), the research of others indicates that learners want their grammar to be corrected (Suphawhat, 1999; Myers, 1997). Even though there has been some opposition to error correction in second language acquisition (SLA) theory (Krashen, 1987) and second language (L2) writing (Truscott, 2004), there is evidence (James, 1998; Schmidt, 2001; Tomasello & Herron, 1989; Chandler, 2003; Bitchener, Young, & Cameron, 2005) that systematically addressing typical errors is conducive to L2 learning. This view is also supported by the fossilised error research described in Pica (1994).

This paper assumes that error correction is beneficial to L2 learners, in particular in the L2 writing process. The cognitivist view of learning is that learners make hypotheses regarding the content studied (Bigge & Shermis, 1999). Swain’s (1998) comprehensible output hypothesis is an SLA theory based on the assumption that learners develop hypotheses about the language they are learning and test them in use. Feedback is exactly what learners need when they are testing such hypotheses. In fact, in real-life communicative situations, L2 learners are often given feedback by native speakers through strategies such as repetitions, confirmation checks, comprehension checks and clarification requests (Long, 1996).

In writing, such assistance is rarely available while writing is in progress, which is why even the long hours invested in writing a paper in L2 do not necessarily make it error free, in particular when the writer shows evidence of fossilized language errors. In order to be effective, feedback must be both timely (Banger-Drowns, Kulik, Kulik, & Morgan, 1991) and appropriate (Levy, 1999). As Cohen and Cavalcanti (1990) point out, teacher written feedback on student errors is not necessarily helpful because it is either inconsistent or not engaging enough. However, conferencing seems to be the students’ preferred channel for receiving feedback, especially on errors (Williams, 2003). In this context, an interactive on-line writing aid, which would respond to the most common learner errors, in a timely fashion, makes a lot of sense.

At first sight, it may seem that grammar checkers, widely available through word-processing software, are the answer. However, they mostly evaluate phrases in probabilistic terms, and thus seem to be designed to support advanced writers, who can more critically evaluate checker feedback. In contrast, apart from potentially confusing the less confident L2 writers, grammar checkers are apparently unable to deal with non-native-like language errors, as they do not meet the requirement of being both tolerant of, and not accepting of errors at the same time. Indeed, a number of ICALL authors pinpoint the latter as a major problem (Tschichold, 2003; James, 1998; Liou, 1991). Tschichold (2003) in particular identified the lack of didactic, semantic, pragmatic and contrastive linguistic knowledge in such parsers as the root of their inadequacy for L2 assistance. Classification of L2 spelling errors by Heift and Rimrott (2008) reveals just how much there is to know about a single aspect of L2 writing. This knowledge is however quite common and often taken for granted in a human teacher (Heift & Schulze, 2007). Although there are on-line writing aids for second language writers, they are according to Meurers (2012) often concordance supported tools which are not necessarily aimed at supporting second language acquisition. Therefore, to address the problem described in the introduction, developing an appropriate, L2 writer specific grammar checker, aimed to support L2 learning, seemed necessary. This development is described in the development section.

Parsers in tutors
Holland, Maisano, Alderks, & Martin (1993) argue that the benefit of parsers, or programs capable of analyzing human language, in CALL is the opportunity it opens for the students to practice not only receptive language skills, but also productive language skills with the help of computer.

The use of parsers in language instruction is commonly referred to as intelligent CALL or ‘ICALL’. It might be more accurately described as parser-based CALL, because its ‘intelligence’ lies in the use of parsing – a technique that enables the computer to encode complex grammatical knowledge such as humans use to assemble sentences, recognize errors and make corrections. (Holland, Maisano, Alderks, & Martin, 1993, p. 28)

There are basically three ways in which a computer can identify and treat a written linguistic error produced by an L2 learner in the target language. It can either perform a pattern matching operation, use a parser or use a hybrid system in which parsing is combined with string matching in an efficient way. Parsing itself can be performed by a variety of parser types. The approach can also vary according to the way the system recognizes and responds to errors. In addition, the system can have a modularity, which allows it to tackle the linguistic levels of the student’s output separately and therefore perhaps more effectively.

According to Matthews & Fox (1991), in order to be able to detect learner errors, an intelligent tutoring system (ITS) or a parser based system must have some sort of a learner language model. While the rules of the correct language constitute the expert language model, the rules of the learner interlanguage represent the learner language model. Matthews & Fox (1991) identify four different principles to choose from when designing the student language model: 1) overlay or rule relaxation (Matthews & Fox, 1991; Menzel & Schröder, 1999), 2) mal-rules or bug rules (Matthews & Fox, 1991; Manning, 1991; Menzel & Schröder, 1999; Heift, 2003), 3) L1 grammar + L2 grammar and 4) robust parsing accompanied by semantic and pragmatic disambiguation. These approaches are explained in more detail in the section which describes the development of ISWT.

The next problem is the frequency of error feedback. While some ICALL authors believe in giving feedback on only one error per utterance at a time, others feel more relaxed about this issue. Pedagogically, too much instructional feedback is thought to overwhelm the student. Instead, it is deemed that specific grammatical phenomena should be focused on from the outset (Heift, 2003). This principle also underlies one of the three recommendations regarding ICALL feedback made by Van der Linden (1993 in Heift, 2003): 1) feedback needs to be accurate, 2) one error message at a time should be displayed and 3) explanations for a particular error should be kept short (not more than three lines at a time). This is in line with the more general feeling about second language error correction, namely that it should focus on a limited number of significant structures, as its objective is learning, and there is a limit to how much a student can learn in one sitting (Mantello, 1997).

Another point to consider is error nomenclature that would make sense to the learner. Granger (2003) bases her error descriptors for feedback purposes on Dulay, Burt and Krashen (1982 in Granger, 2003). These authors suggest two major descriptive error taxonomies, one focusing on linguistic categories (e.g. lexis, auxiliaries, passives etc.) and the other focusing on the way surface structures have been altered (e.g. omission, addition, misformation, misordering). Granger (2003) like James (1998) believes that these two approaches should be blended into one.
Once the error has been diagnosed, the question is whether to correct it or not. Granger (2003), for example, is in favour of providing error correction. According to her, the feedback would be something like “Gender error on pronoun” + correction (L’Haire & Faltin, 2003, p. 489). As pointed out before, there is evidence that error correction benefits L2 learners (James, 1998; Schmidt, 2001; Tomasello & Herron, 1989). Most ICALL authors (e.g. Granger, 2003; L’Haire & Faltin, 2003; Petersen, 2010) in fact believe that what James (1998) calls correction proper should be a part of feedback.

Approaches to feedback style vary. Reeder and Hamburger (1999) believe that error feedback in ICALL should be conversational. They consider it a disadvantage that many ICALL systems respond to learner errors with a template based explanation. Delmonte (2003), on the other hand, takes a different stance. His point of criticism is a feedback taxonomy by Lyster & Ranta (1997 in Delmonte, 2003), which includes explicit correction, recast, clarification request, metalinguistic feedback, elicitation and repetition. These are mostly used by native speakers in conversation with L2 learners (Long, 1996). Delmonte (2003) however believes that anything except explicit correction is inadequate as feedback from the computer. With the advancement of computational language processing techniques, conversational recast entailing feedback is back on the ICALL agenda (Petersen, 2010).

While some of the above proposals are quite general and may not have a specific type of learner at the base, others do not approach the learner from the cognitive psychology perspective (Mantello, 1997). CALL program development however requires careful user and content analysis (Decoo & Colpaert, 1999). Thus, it seems best to make the decisions about the kind of error feedback based on needs analysis, for each learning situation separately. This approach was used in the development of ISWT, to which this article is devoted. The next section describes the development procedure in more detail.

THE DEVELOPMENT OF THE INTELLIGENT TUTOR

Needs analysis

Needs analysis “is a question of major concern” (Oanh, 2007, p. 324) for L2 curriculum development and involves surveying both the target tasks and student performance at these (Benesch, 1996). Two procedures were used to gain an understanding of the students’ learning needs in an understaffed Australian academic English program for L2 student writers. Firstly, a survey of learner styles, means and preferences was developed and administered by a graduate student (Suphawhat, 1999) and secondly, writing samples were collected for the purpose of a learner corpus compilation.

In order to enable both error recognition and correction, ICALL parser developers commonly use learner corpora (Granger, 2003; Dodigovic, 2005; Amaral, Meurers, & Ziai, 2011). Such a parser derives its knowledge of the most common learner errors from a learner corpus (L’Haire & Faltin, 2003; Ott, Ziai, & Meurers, 2012). The corpus can be used either for the development of the parser or to supplement the parser in its use. With ISWT, the corpus was used in the development phase by
researchers who analysed it for specific error patterns, which were subsequently coded into the program.

The study subjects were 87 students enrolled in an academic writing course, mainly speakers of Chinese and Indonesian. The sample also included a small proportion of native speakers of other languages, such as Arabic. Prior to enrolling in the writing course, all students had scored an average of 6.5 or higher on the IELTS test band scale, which is equivalent to approximately 550/580 of the institutional TOEFL scale.

The survey by Suphawhat (1999) indicated that an on-line learning aid, used to complement the work of busy instructors, would be ideal in the situation in which considerable budgetary constraints prevented hiring more tutors. This conclusion was based on the fact that, according to the survey (Suphawhat, 1999), the students seemed computer literate (45.45%) and interested in using the medium computer to improve their English (56.25%). More information on the survey outcomes can be found in Suphawat (1999).

**Learner Style Survey**

The learner style survey uses Willing’s (1988) four-category approach to learner types, piloted and validated in studies by Willing (1988). This approach is derived from a more basic differentiation between two basic groups ‘analytical’ or left-brained and ‘concrete’ or right-brained learners, as established by Witkin, Moore, Goodenough & Cox (1977). The way the analytical learner processes information is linear, sequential and rational. In addition, it is objective, abstract, verbal and mathematical, focusing on detail. Moreover, it is cautious and responsive to selective, low-intensity stimuli (Willing, 1988). In contrast, the concrete learner processes information in a holistic, pattern-seeking, way, which is spatial, intuitive, subjective, concrete, emotional and visual. It focuses on overall impression, while being impulsive and trusting hunches, requiring rich, varied input (Willing, 1988).

Willings’ (1988) paradigm however comprises four learner types. In addition to analytical and concrete learners, each at the opposite end of the scale, Willing (1988) introduces two additional types placed on the continuum line between the two extremes: 1) communicative learners, who are basically analytical learners using communication as a strategy, and 2) authority oriented learners, who are in essence concrete learners with some need for structure.

The results of the survey suggest that most students were analytically oriented in some way, in a very broad sense. The entire questionnaire with the outcomes is the intellectual property of another author and cannot be reproduced here in full. It was however clear from the respondents’ answers to individual questions and question clusters that almost 60% of the students preferred some activities typical of the communicative learning style (learning by conversation), a strong 40% could identify with some of the typical analytical learning style patterns (identifying own errors). It is those 60% that convinced the author to design ISWT as a simulation of the communicative language learning approach, although it is know from the literature (Jokinen & McTear, 2010) that system users do not always prefer a simulation of natural language dialog, but might prefer more “computer-like” output, consisting mainly of parse-trees and short, coded error messages.
In line with the survey result, which clearly favoured a natural communication style, it was decided that ISWT should rely on the human-computer interaction known in the computer world as natural communication (Marsic, Medl & Flanagan, 2000). This kind of interaction emulates human-to-human interaction and is deemed to contribute to the authenticity of a learning situation (Chapelle, 1997; Dickinson, Eom, Lee, & Sachs, 2008). In addition to being conceived of as a Socratic dialog between the student and the computer, the authenticity of interaction is enhanced through a conversational style in approach to error diagnosis and feedback, as advocated by Reeder and Hamburger (1999) and implemented by Petersen (2010).

This means that, contrary to Delmonte (2003), the tutor does not immediately correct the error, but stops to offer the learner some choices, as in line with the findings of Chandler (2003) and Pujola (2001) the learners are not always attentive to excessive feedback and often prefer to self-correct. Therefore, within the framework of ISWT, the learner, having been warned that there is an error, has the choice to self-correct, get a hint or proceed immediately to error correction. This procedure emulates the teacher’s meta-talk, which does take into account the learner's personal style. Thus, as suggested by Dodigovic (2005), it is expected that analytical learners may wish to try out the self-correction route, while the communicative learners may wish to get a hint. This strategy would appeal both to their underlying analytical style and their preference for communication as a strategy. The concrete and authority oriented learners were both assumed to prefer correction; the concrete because correction would be conducive to holistic processing and the authority oriented learners because it asserts the tutor’s authority (Dodigovic, 2005).

While the above explains how the learner style survey was helpful in the development of ISWT user interface, a later section exemplifies in more detail how the interface tackles error diagnosis and correction. The following section discusses the compilation and the use of a learner corpus in the development of the ISWT parser.

**Learner corpus**

A parallel step in the needs assessment was learner corpus compilation. The corpus was analysed for the most common learner errors. This is highly recommended when designing a parser to handle L2 errors (James, 1998; Granger, 2003; L’Haire & Faltin, 2003) and constitutes targeted elicitation of errors specific to a learner population (James, 1998). Such errors are used to design the student language model. In this case, the purpose of the learner corpus was to determine which of the four error handling principles to use: 1) rule relaxation, 2) mal-rules or bug rules 3) L1 grammar + L2 grammar or 4) robust parsing accompanied by semantic and pragmatic disambiguation (Matthews & Fox, 1991).

An overlay or rule relaxation model is based on the notion of “missing conceptions” (Matthews & Fox, 1991, p. 165). This means that the learner’s grammar is considered to be incomplete in relation to the complete target language grammar, i.e. be a subset of it (Yip, 1995). Consequently, the student language model will either omit entire rules or parts of rules. The latter is the case when we relax
some constraints, as the subject-verb agreement in number for example. If this constraint is relaxed, the parser will accept structures such as “*they likes”, in which the subject is plural and the verb is singular.

Another way to present the learner language model is by introducing the erroneous rules applied by the students, which are referred to in the literature as mal-rules (Matthews & Fox, 1991; Manning, 1991; Menzel & Schröder, 1999), bugs or buggy rules or incompetence rules (Matthews & Fox, 1991). Not all types of parsers will of course support this approach. An example of such an incompetence rule is faulty word order (Schwind, 1990 in Matthews & Fox, 1991).

The next solution to the problem of identifying learner errors is presented in the form of combined L1 and L2 grammars. Matthews & Fox (1991) point out that such systems are not technically buggy or faulty since they do not contain any incompetence rules. The underlying theory is therefore that of mother tongue interference. Errors within this framework are not seen as competence errors, since incompetence is not explicitly encoded in the system. They are rather seen as performance errors or misapplication of L1 rules in L2, comparable to the effect of bilingual aphasia (Matthews & Fox, 1991).

A fourth way of dealing with learner errors by way of NLP is based on robust parsing and external disambiguation methods. Robustness refers to the ability to find partial parses in ungrammatical input (Magerman & Weir, 1992). While the first three types of parsers can only process input which matches the underlying morphology and syntax, and will not return a parse for input that only partly matches the rules, a robust parser will return a partial parse for the part which matches any of its rules. This system uses loose syntax rules paired with more constrained semantic and propositional rules. In a way it is a combination of rule relaxation, which happens at the syntactic levels, and introducing extra rules, i.e. by adding new levels of constraints in terms of semantics and domain knowledge.

Each of the above four approaches has some disadvantages. For instance, rule relaxation does not account for all L2 errors, but remains restricted to inflection errors mainly. With buggy rules unanticipated error types may go undetected, as they are not hard-wired into the system, and hence cannot trigger appropriate feedback. In contrast, the L1 + L2 approach restricts all errors to L1 related and drastically narrows down the concept of interlanguage. Finally, robust parsing can require too complex a system of semantic rules. To minimize the disadvantages, it was decided that the choice of the approach had to be based on the kind of errors identified in the L2 corpus.

In order to compile a learner corpus, essays of approximately 600 – 1,000 words in length were collected from the sample of students described above. The corpus of 84,000 words, comparable with other small learner corpora (Pravec, 2004), was then manually analysed and tagged for the types of errors commonly described in the literature and briefly discussed in the theoretical framework section. For flexibility of description, and loosely based on Granger (2003), the error tags were a composite of several markers: 1) level marker (part of speech, phrase, clause), 2) specific variety of level, 3) nature of error, 4) first language of participant. For example, the following sentence from the corpus contains an error (underlined): “It is obviously that a global language is beneficial to the modern world.” In this example, “obviously” was tagged as AdvAdjIndon, indicating that the error occurred at the part of
speech level, more specifically adverb (Adv), which was mistakenly substituted for adjective (Adj). The Indon tag component marks the first language of the participant as Indonesian.

Error ranking according to type and frequency followed, indicating that approximately one third of all errors were related to verb. Upon further analysis, these were found to include pseudo-passives, ergative construction, tough movement and existential construction (Yip, 1995) as well as missing copula and finite – non-finite verb confusion (Yong, 2001). Within each part of speech category, part of speech confusion was the most frequent error type (Yong, 2001). However, other errors were identified too, some of which had to do with inflection and hence required rule relaxation. While the buggy rule approach was selected, some rule relaxation was allowed as well.

One unexpected outcome of the corpus analysis, which also kept track of the first language of the participant for each identified error, concerned the seven major structural errors and a few minor morphological ones. While four of these major errors were originally assumed to be specific to Chinese-English Interlanguage (Yip, 1995), and the remaining three to Indonesian-English Interlanguage (Yong, 2001), error frequency by first language showed that all of the seven major errors were made by learners across the board, regardless of their L1.

How ISWT works

The backdrop for the interaction between the student and ISWT is the topic of malaria, on which the student is supposed to write an expository essay. ISWT asks Socratic questions about the topic and the student is supposed to enter free-style answers to these questions. The sequence of the questions provides the scaffolding for the organisation of the essay. The questions are also designed to encourage summarizing and paraphrasing of the main ideas rather than plagiarism. Some form of plagiarism is often found in the L2 writing of students who lack confidence in their own language ability (Amsberry, 2010; Erkaya, 2009; Ballard & Clanchy, 1984; Simmons & Thurstun, 1995).

All student input is parsed. Lexically, orthographically and grammatically correct answers receive praise along with a tree diagram representing the structure of the sentence (Figure 4). According to Gregg (2001), the use of linguistic metalanguage, as found in the parse tree, constitutes a type of evidence with respect to the learner’s hypothesis about the target language, which according to James (1998) leads to an understanding of not only what is correct but also why it is correct. Incorrect answers are handled differently. After diagnosing an error, ISWT gives the user three options: “try again”, “get a hint” and “see the solution” (Figure 1). The hint is based on both the error type and learner input.

The “try again” and the “get a hint” (Figure 2) options are thought to appeal to analytical and communicative learners respectively, since educational psychology describes them as learners who like to solve problems in a step-by-step approach that might benefit from opportunities to work out a problem without assistance or from a partial solution (Willing, 1988; Witkin, Moore, Goodenough & Cox, 1977). In light of the same theory, the “solution” (Figure 3), is designed for concrete and authority oriented learners, who are deemed to be right-brained and therefore might require experiential learning (Willing, 1988; Witkin, Moore, Goodenough & Cox, 1977). The intended outcome of correction in this context is that the learner would gain experience with model target
language, which would, according to the nativist view of L2 acquisition (Gregg, 2001) provide some positive evidence deemed crucial to learning. In light of interactionist L2 acquisition theory (Long & Robinson, 1998; Doughty, 2001), the correction is a recast (Spada & Lightbown, 1993), or the least disruptive cognitive intrusion likely to have the desired repair effect.

Figures 1 – 4 below illustrate the interaction between ISWT and a student, limited to answering one system question. In Figure 1, question 4 sets the following paraphrase task: “One of the reasons why malaria is so difficult to control is the occurrence of new factors. How would you introduce the occurrence of a new problem, e.g. in your essay?”. By answering “There is a new problem occur”, the student commits the existential construction error, which is why the student input is flagged red and accompanied by feedback “Sorry, that was incorrect”. The options “try again”, “get a hint” and “see solution” are offered. In Figure 2, the student, having requested a hint, gets the following ISWT suggestion: “Try completing the phrase ‘There is … or … has occurred’”. The student then requests a solution, which is provided in Figure 3: “There is a new problem or A new problem has occurred”. The “continue” button shown in Figure 3 gives the student the opportunity to self-correct, based on the ISWT recast. This option was given based on the understanding that adult learners might be likely to engage in repair-uptake of the correction (Panova & Lyster, 2002). Of course, ISWT clears the screen, if the student chooses to continue. In Figure 4, the correction-uptake has already been processed by the system, generating a parse tree and praise.

![Figure 1: Existential construction detected](image)
Dialogue about an essay topic - lesson 1

Question 4

One of the reasons why malaria is so difficult to control is the constant occurrence of new factors. How would you introduce the occurrence of a new problem, e.g., in your essay?

Your (correct) input was: There is a new factor occur

Good solutions would be:

- There is a new problem
- A new problem has occurred

Continue

Back to lesson menu

Figure 2: A hint is provided at request (based on input in Figure 1)

Dialogue about an essay topic - lesson 1

Question 4

One of the reasons why malaria is so difficult to control is the constant occurrence of new factors. How would you introduce the occurrence of a new problem, e.g., in your essay?

Your (incorrect) input was: There is a new factor occur

Figure 3: Solution is provided at request
As can be seen from the above examples, ISWT does not provide immediate explicit correction, at least not without consulting the learner first. It does not provide metalinguistics error description either. Its feedback delivery is rather conversational, as advocated by Reeder and Hamburger (1999) and implemented by Petersen (2010). Based on needs analysis, as recommended by (Decoo & Colpaert, 1999), ISWT gives learners a choice of correction level, so that each learner type can get an appropriate kind of feedback. This can be a corrective hint, a full correction or the opportunity to self-correct. The language rule is delivered indirectly, through the full parse of the correct sentence, but the learner must first enter the correct solution to see this. This step may not only be a part of the self-correction process, but could follow the uptake (Panova & Lyster, 2002) of the corrective feedback and facilitate learning.

**Summative evaluation**

Software evaluation is often divided into two procedures: formative evaluation conducted during the development process and summative evaluation conducted at the end of the development process (Gediga, Hamborg, & Dutch, 1999). While the formative evaluation of ISWT was conducted during the development phase, by having a representative sample of students and instructors use the software and provide feedback, the summative evaluation followed the completion of the software prototype. This study aimed at establishing the level of effectiveness of ISWT, the level to which it can be
successfully utilised by L2 academic writers to improve their English, especially in situations with no or limited language instruction. Thus, it became imperative to establish whether or not the use of ISWT promoted learning. Accordingly, the goal became to collect quantitative performance data or scores (Goodfellow, 1999).

Since the purpose of ISWT was to provide feedback on grammar mainly in self-study mode, there was no point in comparing the outcome with teacher induced learning. In more recent literature, the two are considered incompatible and comparisons between a teacher and a computer program are discouraged (Goodfellow, 1999; MacWhinney, 2001). Moreover, the premise on which the study was constructed was that the treatment should address what is known as fossilised language errors (Selinker, 1972; Pica, 1994; James, 1998; Yip, 1995). Such errors, as pointed out earlier on, emerge at a certain stage of L2 development and persist despite continuous language instruction. Based on the learner corpus analysis, this seemed to be the case with the sample on which the needs analysis was conducted. The fossilization of the seven most prominent error categories, described earlier in this text, was evident in the persistence of such errors through multiple writing assignments. Although the needs analysis participants were all enrolled in an academic English class, their fossilized errors seemed to persist.

Consequently, the summative evaluation study included elements of within-sample design (Al-Seghayer, 2001; Felix, 2008), which was combined with limited comparison. The two participating groups and their roles are explained in further text. In the course of one semester, during which all the treatment group participants were enrolled in an academic English class, three written assignment samples were collected from each participant to establish the existence of the most prominent types of fossilized errors. The treatment was introduced at the end of the semester, and only administered to such students in whose writing repeated errors were identified.

To capture a sample similar to the one used in the process of needs analysis, the subjects were taken from three different populations (Taiwan, Australia and UAE). The subjects in Taiwan were all speakers of Chinese, the ones in the UAE speakers of Arabic and the ones in Australia speakers of Chinese and Indonesian mainly. There was no randomization or pair matching, as this was a convenience sample. On the other hand, matching with the software designer’s objectives in a number of confounding variables, such as age, English proficiency level, fossilised L2 errors, education profile and first language, provided some controls and counterbalances (McDonough & McDonough, 1997). These variables closely corresponded to those identified in the needs analysis study. Thus the subjects were all university students or applicants, aged 19-21, with paper-based TOEFL score of approximately 500-550. In addition, close observation of the participants’ writing over a semester indicated the repeated presence of fossilised language errors such as the ones identified in the needs analysis corpus.

The treatment in this study was preceded by a pre-test and followed by a post-test (McDonough & McDonough, 1997; Larsen-Freeman & Long, 1991). The pre-test was a grammaticality judgment test (Ellis, 1997). In this test, the takers must literally judge the grammaticality of utterances. The utterances were based on the most frequent errors identified in the learner corpus. This kind of test is helpful since certain aspects of grammatical knowledge cannot be understood by mere analysis of production data (Yip, 1995), such as the three samples of each participant's writing used to select...
study participants, because frequent avoidance of difficult structures in L2 production (Yip, 1995) can give a misleading account of the learner errors. This partly justifies the choice of this instrument in the study. An additional reason for selecting the grammaticality judgment test was that it was likely to reveal real errors of competence, rather than slips of pen or mere mistakes of performance (Ellis, 1997; James, 1998).

There were twelve questions in the grammaticality judgment pre-test. Each question was very similar to those asked by ISWT. The task in each question was to identify one or more correct paraphrases of a statement. The erroneous distracters in this test contained, but were not restricted to, the seven most common errors, all of which were, as already pointed out, identified as significant in both the learner corpus and the participants’ writing samples. The results of the test demonstrated that four errors were consistently made by the students: tough movement (TM), part of speech confusion (PSC), finite – non-finite verb confusion (FNV) and ergative construction (ERC).

The post-test, administered within 24 hours of the pre-test, and 22 hours from the completion of treatment, was a short answer test, which allowed the students to produce their own sentences. The format was deliberately different to the format of the pre-test in order to prevent the mere learning from the pre-test to influence the outcome of the post-test (Ellis, 1997). Larsen-Freeman & Long (1991, p. 32) successfully address concerns one might have regarding test task incongruence by reporting on several studies that found no significant differences between the rate and type of errors elicited through a variety of tasks. Improvement was measured in terms of the amount of error found in the student-written sentences. While the anticipated errors were the focus of this instrument, all errors made by the students were recorded, even if they were not the anticipated ones. However, few unanticipated errors actually occurred.

To address the concerns arising from the difference in error elicitation tasks between the pre-test and the post-test and in order to counter any measurement error that could potentially be introduced through the measurement instrument, the pre-test and post-test were piloted on a representative sample of 59 takers, similar to but not included in the treatment group (Hughes, 2008). The results, which are discussed in more detail below, indicated that the pilot group takers did not perform better on the post-test than on the pre-test. Limited though in scope, the outcome of test piloting would seem to suggest that the instrument was not likely introduce a statistically significant error in measurement.

The treatment was the work with ISWT (McDonough & McDonough, 1997). The students at each of the participating institutions worked simultaneously on the trial in the computer lab of their respective institution, under the supervision of two teachers at a time, in block periods of 120 minutes each, after completing the pre-test and before taking the post-test. At each of the institutions, the treatment occurred between 10am and 12pm local time. The first hour was spent reading several texts on the topic of malaria, which reduced the actual work with the software to approximately 60 minutes. The students worked in self-study mode, but were not prevented from consulting with their classmates. However, no consultation was allowed during the pre-test and post-test phases. The software was introduced to the students and troubleshooting assistance was available during the trial. Overall, 266 subjects participated in this study. Of these, 107 were in the United Arab Emirates (UAE) (60 in the Emirate of Sharjah, 47 in the Emirate of Dubai), 83 in Australia, and 77 in Taiwan.
Summative evaluation - Results

The post-test showed an average reduction in error rate of 83% with respect to the pre-test across the entire sample. As shown in Table 1, the highest score was achieved by the Taiwanese students (94% error reduction rate), followed by AUS students (85% error reduction rate), ZU students (79% error reduction rate) and then the overseas English language learners in Australia (73% error reduction rate).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Error reduction rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taiwan</td>
<td>94%</td>
</tr>
<tr>
<td>UAE Institution 1</td>
<td>85%</td>
</tr>
<tr>
<td>UAE Institution 2</td>
<td>79%</td>
</tr>
<tr>
<td>Australia</td>
<td>73%</td>
</tr>
<tr>
<td>Average</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 1: Error reduction rate

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Test Piloting Group</th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.88</td>
<td>0.31</td>
<td>Mean</td>
<td>1.69</td>
<td>1.81</td>
</tr>
<tr>
<td>SD</td>
<td>1.19</td>
<td>0.62</td>
<td>SD</td>
<td>1.12</td>
<td>1.14</td>
</tr>
<tr>
<td>SEM</td>
<td>0.07</td>
<td>0.04</td>
<td>SEM</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>N</td>
<td>267</td>
<td>267</td>
<td>N</td>
<td>59</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 2: T-test data summary for treatment and test piloting groups

A related samples t-test was conducted and the reduction in error rate was found to be statistically significant (t = 19.8209, p < 0.0001). By contrast, the pre-test – post-test piloting procedure, involving the administration of the two tests to a similar profile sample of 59 within the same time span, but without the intervention in-between, was also subjected to t-test analysis, yielding a difference in results which was not statistically significant (t = 1.2237, p = 0.2260). In fact, the error rate increased slightly (by approximately 6%), but not statistically significantly, indicating that learning from the pre-test administration was not very likely. T-test data summary is found in Table 2, consisting of mean error rate (Mean), standard deviation (SD), standard error of measurement (SEM) and the group size (N).

Summative evaluation – Discussion and limitations
While it is true that the pre-test and post-test piloting group was smaller than the treatment group, it showed no significant difference with respect to the treatment group on the pre-test ($t = 1.1148$, $p < 0.2658$). Moreover, it was very similar to the treatment group in demographics. Three samples of academic writing were collected from both groups and analysed for the presence of the typical errors identified in the initial student corpus. These samples were coursework written over one semester by each of the participants. The nature and frequency of errors identified in the samples coincided with those found in the initial student writing corpus. The error persisted, although all of the participants had 3 – 5 hours of writing instruction per week. Prior to the semester during which the work was written, the participants were exposed to 6 – 12 years of English instruction. Many of them seemed genuinely interested in achieving an improvement in their English. In other words, the participants are likely to have tried and failed, numerous times, to eradicate such errors in their writing. Consequently, it could be argued that the student track-record in writing provides some within-sample evidence of the persistence of the fossilised errors. Therefore, it would be less likely that some random event, other than exposure to ISWT, would have caused a comparable drop in error rate.

An obvious question to ask would be whether the errors were simply avoided in the post-test, which was open-ended, by choosing alternative grammatical structures. This however was not the case. The students attempted the same structures they were struggling with in the previous two sessions, the pre-test and ISWT respectively. However, this time their rate of success was much higher. Another possibility is that the students might have memorized entire sentences while working with pre-test or ISWT. Their post-test responses were however mostly different in many ways from the example sentences presented in the pre-test and ISWT. Finally, the likelihood that the students copied from each other was small, as the answers differed from student to student. Thus, the results seem to suggest a genuine improvement in target grammar.

Another concern is the fact that there are similarities and overlaps between the Socratic prompts generated by the software and the post-test questions. The answer to this concern is found in the principles of language assessment. According to the basic requirements in test design (e.g. Stoynoff & Chapelle, 2005), one aspect of overall test validity is content-related evidence of validity. In particular, a curriculum-based test should reflect the content it is supposed to measure. Since the software in question covers a very small curriculum and is introduced to the study subjects once, over a very short period of time, the most obvious test design to measure its success would be an achievement test (e.g. Brown, 2004). Such a test should assess what is covered by the given curriculum. With a very small curriculum, as is the case here, the post-test questions are bound to be very similar to the content presented by the software.

Several other factors might have contributed to the improvement in the participants’ grammar, as recorded in this study. The first one is called the ‘Hawthorne effect’ (McDonough & McDonough, 1997, p. 166). This effect refers to a substantial improvement in student performance under study conditions due to their noticing the uniqueness of the situation. While the quasi-experimental procedure was introduced in a fairly low-key manner, it differed from a usual class in the location (computer lab) and the presence of the researcher. Since the entire procedure was conducted by this
author, who is also the author of the ISWT, it is of course possible that the author’s enthusiasm for the software (McDonough & McDonough, 1997) contributed to the success of the study subjects.

Finally, the pre-test itself might have contributed to learning, thus enhancing the effect of the treatment. Ellis (1997) reports that some tasks can significantly raise the learners’ consciousness concerning some linguistic property of the target language. Judgment of well-formed vs. deviant linguistic data, or in other words our grammaticality judgment test, is precisely such a task that could have helped a learner arrive at an explicit understanding of the linguistic item in question. However, no studies come to mind where a statistically significant difference between the pre-test and post-test was completely unrelated to the treatment itself. Besides, the results of the pre-test and post-test piloting indicate that learning from pre-test would not have been likely. In addition, language instruction could be ruled out as a possible cause of improvement, since prior to the treatment, the fossilised errors persisted in the participants’ writing throughout an entire semester of studying academic English. Therefore, a statistically significant improvement in the participants’ performance seems to suggest that the ISWT program has been able to serve its purpose.

No qualitative data has been collected in the course of this development study. Any future work on this software prototype would benefit from data obtained from surveys, interviews and think aloud protocols. In particular, it would be interesting to see which feedback option participants would select and how it relates to their cognitive style.

CONCLUSION

This paper follows the development of an intelligent tutor, from needs analysis to a trial implementation of the prototype. The purpose of the software has been to give intelligent feedback triggered by sentence grammar errors in L2 English. Needs analysis conducted with the aim of developing this software included both cognitive styles and typical errors of non-English speaking background students studying at an English medium university. While the research informed error typology was used to develop the parser, i.e. the main language processor of the program, the cognitive style information helped shape the user interface. Since the original student sample was no longer available by the completion of the development, the prototype was tested using a sample with very similar characteristics. In a summative evaluation study, the difference between the pre-test and post-test data was found to be statistically significant, suggesting that this software contributes to learning and could thus be used in educational situations, where access to English instructors is limited or impossible. Further examination of student-computer interaction, based on the cognitive style of students, would be desirable.

This paper contributes to knowledge in several ways. Firstly, it brings together a number of ideas and theories across disciplines of second language (L2) acquisition, Intelligent Computer assisted Language Learning (ICALL), second language teaching, education, needs analysis and software design, in an effort to demonstrate the usefulness of the combination. It also exemplifies L2 error correction at work, in a unique setting. In doing so, it additionally discovers new facts about the nature of fossilized L2 errors.
ACKNOWLEDGEMENTS

The research and development work described in this article was funded through several Macquarie University grants, a portion of a Zayed University research grant and a seeding grant awarded by the American University of Sharjah. A debt of gratitude is owed to Alison Fowler for contributing her computational expertise in the development phase. Finally, I am indebted to the anonymous referees for their thoughtful comments and encouragement.

REFERENCES


