

Tagging the interlanguage of Chinese learners of English

John Milton and Nandini Chowdhury
(*Hong Kong University of Science and Technology*)

Introduction

This study reports on a newly emerging application of the techniques of corpus linguistics: the analysis of corpora of interlanguage. Here, we describe the initial stages of the collection and tagging of a corpus of the written interlanguage of Chinese learners of English. While there are differences between this and other studies of interlanguage corpora (e.g. Granger et al., this volume) in collection techniques, size of corpora, learner varieties, proficiency levels and approaches to tagging, together they provide a novel approach to the analysis of learners' writing. More importantly, they furnish data about student performance that are crucial for the development of new pedagogical instruments. This study has as its ultimate goal the creation of automatic grammar and writing tutorials that will address some of the most persistent difficulties non-native writers have in dealing with English syntax, lexis and semantics.

In compiling the interlanguage corpus for our project we began by transcribing 1,100 scripts (545,900 tokens) from the 1992 Hong Kong public matriculation examination — about 6% of the 16,103 composition scripts of students sitting the examination that year. The Hong Kong Use of English Examination is taken each year by students leaving secondary school and is the placement instrument for tertiary English-language programmes. Scripts were selected from two categories of grades assigned to the compositions by the Hong Kong Examinations Authority: 'D' and 'E' (within each of these grade ranges the scripts were randomly chosen). The examination scripts are statistically representative of the larger population from which they have been culled. The 550 scripts we have from each grade range represent about 10% of those in each category.

We chose first to examine the writing of students at this level of performance because it is this cohort of students¹ which has generally been identified by tertiary institutions in Hong Kong as most in need of English language tuition. For example, students at the Hong Kong University of Science and Technology with grades of D and E in this examination, are mandated to enrol in the first-year English Language Enhancement course (this group consists of between 70% and 80% of all students entering this institution). Students with grades below E in this examination are generally not admitted to the University. The reliability of this examination in discriminating for university placement is well documented (cf. Lewkowicz et al. 1991).

This examination-derived corpus is divided by topic: students had the choice of writing on four expository topics. We have, up to now, manually tagged a random sample of about 8% (77 scripts, consisting of 42,456 tokens) of the corpus. This annotation has proceeded at the same time as the keyboarding of the corpus from the students' handwritten scripts.² We are also compiling an archive³ of students' word-processed assignments submitted for the first-year English course. This archive now stands at close to 5,000,000 words, arranged by topic. We are currently developing a representative corpus from the archive of the students' untimed writing. Both examination and assignment corpora will be analysed for variations between the writing students produce under examination conditions and out-of-class assignments.

The issues that can be explored through this study of interlanguage variation should help us better understand how students perform in various writing circumstances and should allow us better to assist them to write more effectively under these conditions. For example, preliminary comparison of the two corpora suggests that the lexical densities of the untimed assignments and examination scripts are, as might be expected, very different. Surprisingly, however, the range of vocabulary (measured by the type/token ratio) employed by students is much higher in the examination scripts than it is in the untimed assignments. This may be because students memorise expressions and vocabulary in preparation for the examination, that are then abandoned in the course of the summer months before they enter university.

We are continuing to enlarge the size of our corpora and broaden their scope, in order to increase their representative value as models of the range of the written interlanguage of Hong Kong students. Opportunity has allowed us to be systematic in our collection techniques; both English examination scripts and assignments from language courses are readily available. These are likely to be the only sources of corpus materials, as English is becoming an 'examination language' in Hong Kong.⁴ We also hope to transcribe examination scripts from other grade ranges. We should soon have compiled corpora, a significant amount of it tagged, that reliably represent the range of students' written English in both scale and type.

Annotation

These archives have already, in their untagged form, provided grist for some speculation on the source of error in the use of connectives in students' writing (cf. Milton & Tsang 1993). However, while considerable research can be conducted with the interlanguage corpus in its untagged form, the morphologically ambiguous nature of English, compounded by the particular ambiguities of the students' writing, dictates that much more data about the patterns of this interlanguage can be retrieved from an annotated corpus. Our approach departs, as far as we can

determine, from previous error analysis studies by attempting large-scale tagging of all lexical expressions in the corpus. This will enable us to gauge relative frequencies of error in measuring interlanguage variation and in comparing the interlanguage to its target. We hope in this way to be able to address issues such as whether the frequency of error in relation to non-error represents genuine difficulty for the students, how the variables of writing circumstances affect learners' writing, and the degree to which error probability can be measured in order to produce automatic tagging algorithms.

Although corpus linguistics attempts to operate from a basis of quantitative empiricism, much of any annotation scheme, automatic or manual, must be introspective and rest on native-speaker intuition. In the annotation of interlanguage, this introspection operates on at least two levels: assigning tags based on theories of how the target language operates (cf. Sinclair 1991b:39) and attribution of 'error' in the interlanguage. The present discussion addresses the latter issue.

An obvious problem is that of accounting for the uncertainty of error type. We attempt, wherever there is insufficient evidence to assign one interpretation, to indicate alternative possibilities. For example, a frequent problem our students have is with the misuse of articles, and the resulting confusion between singular and plural forms, as in the following sentence:

They regard examination as if it were a hell.

The error in this sentence may lie in the student's failure to add -s to the word *examination*, or in omitting the definite article before the word. The presence of the anaphoric *it* is no guarantee that the student intended the object to be singular. If we indicate that either interpretation is possible, we must also have a way of noting that *it* is an error if *examination* is plural. This process of accounting for various possible errors, including those which are only apparent when antecedent corrections are made, is limited by the patience and imagination of the investigators. While we try to annotate reasonable alternative interpretations, it is doubtful that any analysis could guarantee total accountability of every possible option. Tagging a learner corpus allows us, at least and at most, to systematise our intuitions.

Purpose

In addition to providing empirical information about the state of the writing of Hong Kong students, this large-scale analysis of interlanguage will provide information for the creation of pedagogical aids such as electronic composition and grammar tutorials directed to the needs of these students. Other projects aimed at developing second-

language grammar checkers have been based on the analyses of relatively small interlanguage corpora, and tagging has been limited to the annotation of error only, with no attempt to describe how often students use the same structures correctly. Also, these projects have sought to emulate the same corrective techniques of current commercial programs (cf. Liou 1992). A fully representative and systematically tagged model of the writing of Hong Kong learners of English will form the basis for more effective automatic error analysis than is possible with our current knowledge of this interlanguage. Knowledge of the students' linguistic and pedagogical needs must be incorporated into computer programs designed to provide some of the realistic and practical advice to students that is currently not available except from a human expert.

Why, it might be asked, since the errors that particular groups of learners make are generally well documented, is it necessary to study the interlanguage in such minutiae? Why not just dump all recognised errors into a database and have the program flag every matched occurrence in students' writing? The answer is that while lists of the common errors of Chinese students are indeed available, we have no information on the frequency of these errors among this cohort of students. Neither do we know the context in which these errors are likely to be made. Without this information, we cannot develop a program that would operate on a rule-based system (such as probabilistic algorithms) to identify errors. Computational issues should also be considered: it is unlikely that pattern-matching techniques alone could reliably flag every occurrence of a particular error-type.

Our focus on a practical pedagogical purpose for this study is made compelling by several developments. One is the increasing number of students entering higher education in Hong Kong, whose English is not commensurate with the demands of tertiary education.⁵ A second is a technological development that gives some urgency to the provision of empirically-based and pedagogically sound electronic writing aids to our students. This is the proliferation — and popularity — of commercial 'grammar checking' programs, the advice from which is inappropriate to our students' needs. Our students' use of grammar checking programs that are now available with most word-processors (e.g. *WordPerfect* and *Microsoft Word*) may actually result in a deterioration of their writing (cf. Pennington 1991).

It seems reasonably unassuming to claim that the quantitative description of the written language of our students will be of advantage in designing facilities for the improvement of their written expression. One of the pitfalls we intend to avoid is that of trying to give more information than the program and our current state of knowledge can reliably provide, and thereby frustrating the credibility of our procedures. We also mean the interactive tutorials we develop with this information to go beyond mere anodyne grammar correction. For a short description of a suggested design of a writing tutorial using the macro programming of word-processors, see Milton (1993).

Method

Automatic annotation is currently reliable in the form of word-class tagging. We have manually tagged error mainly at a relatively 'surface' level in order, at a later stage, to test how effectively an automatic tagger can detect error. It is also computationally much simpler to associate error with form-function relationships at the word and collocation level than with larger or more amorphous linguistic units and relationships. We are not, after all, accounting for rhetorical error. We do, however, indicate inter-sentential and word-order problems when there is no more discrete way to account for error.

The only tagging program produced specifically for the annotation of interlanguage that we know of is *COALA*, still undergoing development at the University of Sydney (Pienemann 1991). This program appears to provide the machinery for the assignment, storage and retrieval of clause parsing, but it does not contribute to word-class disambiguation. In the absence of a program that provides such assistance or a tagset specific to interlanguage, we have rigged our own tagging mechanism by mapping a keyboard with our own tagset.

Given various constraints, such as the dearth of work in processing interlanguage for computer analysis, or a conventional tagging system for interlanguage, we were faced with some very basic questions in the choice of our tags. The pragmatic nature of our purpose dictates, at least initially, that we label the interlanguage based on patterns that arise within the texts rather than waiting for some ultimate refinement of a standardised tagging system for NS corpora, and then modifying this.

The proof of the effectiveness of our, admittedly *ad hoc*, tags will be whether they provide us with a retrieval mechanism which will afford a more complete view of the interlanguage than we now have and, ultimately, an opportunity for a degree of automatic diagnosis of error and prescription of remediation.

Parameters

Our tagset, and therefore our categories of error and non-error, have grown out of the patterns that emerged as we proceeded through the corpus. The following are some of the questions we have addressed in determining our current tags.

To what degree should the tagging system reflect a constituent-structure hierarchy and a hierarchy of delicacy of detail? Such hierarchical issues include: the priority of syntax labelling, e.g. by word-class category; constituent structure and grammatical function; the recognition and definition of collocational boundaries by the assigning of unified tags to more than one graphic word, i.e. words which form a cohesive unit; and the isolation of suffixes and prefixes from the base form of words.

*Should tags distinguish subcategories within a word class, according to such differences in their function as co-ordinating vs. subordinating conjunctions, superlative vs. comparative adjectives and demonstrative vs. reflexive pronouns? Most automatic tagging systems do, but working as we are through the corpus manually, we limit our detail to distinguishing between morphologically identical subclasses (e.g. existential *it* vs. referential *it*).*

If in the analysis of interlanguage we are working under a different set of constraints from those of the target language, how do we interpret and analyse ambiguous structures? Even the syntactic analysis of written and spoken discourse in a first language is partly intuitive, as researchers have to decide how to interpret sentences based on what may be the disparate semantic, syntactic and world knowledge of the speaker/writer and the researcher. This problem is magnified in the analysis of interlanguage, especially in this particular interlanguage where the intended meaning is often not clear. Two related forms of ambiguity arise: one where the meaning is not clear and another where the meaning is at least apparent, but where more than one reconstruction is possible.

To what extent do we need to base our structural analysis at this stage on semantic and discourse structure? If we must consider syntagmatic (chain) and paradigmatic (choice) relationships, issues of both structure and meaning among linguistic constituents, boundaries need to be established.

How should punctuation be tagged? While we do not tag unique symbols (which are easily retrieved by the nature of their uniqueness), it does seem advantageous to tag such common punctuation errors as comma splices.

Our efforts to make tags carry as much information as possible have resulted in sometimes intricate annotation, but the apparent cumbersomeness of the tags does not hamper our analysis procedures. Our primary purpose has been to adopt a tagset which will provide useful indices for retrieval. Admittedly, describing alternative possibilities for error in practice makes for complicated labels, but a string search on these labels is straightforward. The availability of powerful text-retrieval software means that any text, whether tagged, indexed or not, can be subjected to the same search and retrieval mechanisms as any structured database.

In order to formulate valid indices we have attempted to be systematic in our use of grammatical authority, although we are less concerned with the theoretical pedigree of our grammar than we are with its consistency and inclusiveness in describing this interlanguage. To this degree, our tags are not entirely *ad hoc*, as they follow the consensual categories of other theory-neutral tagsets.

Phrasal, clausal and word-class boundaries

One danger of tagging a large corpus of interlanguage is the risk of inconsistency in interpretation. This is another reason why we have mainly applied low-level rules (i.e. word-class assignment guidelines and, wherever necessary, morphological characteristics): to be as reliable and consistent as possible in the initial tagging phase. Nevertheless, we have found it practicable to tag for some categorial constituents since their functions assist in identifying relations between constituents and word-order rules.

If complete parsing were attempted, including the interpretation of discontinuous constituents, sentence fragments, comma splices etc., the annotation scheme would become very unwieldy (not to mention that automatic parsing is not yet possible, so a complete parse of the corpus is not practicable). By limiting analysis largely to the word-class level we ensure that each word or collocational unit has at least one tag marking the grammatical category and, where necessary, a tag denoting approximate error-type.

Collocations

Although tags are normally attributed on the basis of single-word syntax, we find it useful to tag certain syntactic units and collocations at the phrasal or clausal level, e.g.

- (a) proper multiple-word nouns (*Hong_Kong_University_of_Science_and_Technology* {*npr*});
- (b) *to* + infinitive (*inf*);
- (c) fixed syntagms that function as single grammatical units such as quantifiers (*many_of*{*quan*}) and prepositions (*according_to*{*prep*});
- (d) hyphenated compound adjectives (*newly-developed*{*adj*}) and other similar coinages;
- (e) phrasal verbs and word groups that function as conjunctions (*such_as*{*conj*}, *in_order_to*{*conj*}); and
- (f) phrases such as *at_length*{*advp*} and *and_so_on*{*conj*},

where the collective sense is somewhat different from the sense of individual word-forms.

By choosing to tag word-sets collocated by the students, we sacrifice the syntactic identity of single words in the expression. Collocational annotation allows us to expose patterns unique to this interlanguage. Native English collocations are tagged at the phrasal level (e.g. *pillars_of_society*{*np*}). Because collocations appear to be frequent focuses of errors, it seems justifiable to tag at the phrasal level at the expense of word units that are less likely to be problematic. Commonly occurring collocations which are distinct from those in the target language are tagged so as to be retrieved as one segment. For example, collocations of the preposition *of* to signal association/possession (*the_age_of_knowledge*, *the_age_of_growth*, *the_age_of_pursuing_knowledge*) are tagged to indicate idiomatic error (where {*np#i*} is used to signify a noun phrase with an error in idiom).

Certain units used invariantly by the students in this corpus are collocated (e.g. *lack_of* is used by students in place of the verb form). Conspicuous transliterations of native idioms (*a_needle_does_not_have_two_sharp_points*) are also colligated and tagged as unique expressions ({*exp*}). Focusing subjuncts such as *not_only ... but_also* and other similar connectives such as *so ... that* are colligated using the two-part tag {*advp1*} and {*advp2*}. In the case of connectives such as *as ... as* we use the tag {*advp1*} and {*advp2*} to highlight the connection between the two. Contractions such as *can't* are tagged as {*mod-adv*}. Deviant collocations are marked using the tag {*#i*} ('unidiomatic collocation'). In cases of word misuse within compound units (which are tagged as single lexical items), e.g. when *love* is used instead of *love-affairs*, we use the tag {*np#+/-*} to indicate that one word within the compound unit is either redundant or missing.

In general, the broad criteria for tagging units as collocations (despite the admittedly fuzzy boundaries) are: instances where the presence of a unit is the result of a single choice, e.g. *according_to*, *such_as*, *and_so_on*, *not_only ... but_also* etc.; and expressions and idioms used either invariantly or with only occasional internal syntactic/lexical variation.

Colligating those items which co-occur frequently and form distinct lexical units helps determine the contexts in which learners use target language collocations; how frequently they are used; what deviations occur; and what functions these collocations have, as compared to NS usage.

Morphological boundaries

For our purposes we do not need to use separate tags marking affixation for all occurrences. However, as our tagging process is context-sensitive (i.e. tags are assigned depending on the linguistic environment), we need to examine structures based on form-function relationships.

When ambiguous structures relating to omission, substitution or overgeneralisation of affixes occur, it is advantageous to isolate morphemes. Errors relating to singular and plural noun forms, agreement and verb tense are exceptions and are tagged separately without isolating the morphemes. An example of omission is:

It{expro} is{vps} very{adv} important{adj} for{prep} you{pro} to_study{inf} the{art} most{quan}
advance{adj#+d} technology{nu}

In this sentence *advance* has an adjectival function, but a noun form. We therefore tag the word as an adjective-missing-a-morpheme. Affixes are often similarly isolated in compound units, e.g.

In{prep} recent{adj} years{npl} it{expro} is{aux*w} become{vps} a{art} <quan>
talk_of{adj#+ed#prep} issue{ns} in{prep} school{ns}

In this case the first unit of the correct form, *talked-about*, is truncated and the preposition is wrong.

Overgeneralisations of the *-ed* participular ending are tagged using the symbol *-ed*, e.g.

new{adj} technology{nu} that{conj} **costed{vpas#-ed}** large{adj} sums{npl} of{prep}
 money{nu}

However, in instances where students substitute morphemes, or coin words, we use the | sign to signal substitution. We also recognise ambiguous error of this type by offering an alternative spelling error tag, e.g.

Hong_Kong{npr} is{vps} a{art} **pragmaticive{adj# | /#sp}** society{ns}

The tag for spelling error is used to suggest the possibility that the deviation may be a random slip, whereas the symbol | suggests that morpheme substitution may be consistent. The corpus offers evidence of frequent misuse of the *-d* and *-ed* increments, and thus these do not appear to be random spelling errors.

The following example highlights the problem of identifying form-function relationships in a developing system. Here we have an instance of deviance in affixation, negation and collocation, all at once:

And{conj} it{pro} will{mod} make{vbase} the{art} students{npl} **can't_concern**
{mod~adv_adj#+ed#neg |} about{prep} their{padj} school{nadj} work{nu}

The tag signifies a collocation of contracted modal (*can*) + adverb (*not*) and adjective missing *-ed* (*concern*), with an error in negation substitution.

A common feature of the corpus is conjoined words (marked by #_), which in NS English occur separately, e.g.

They{pro} cannot{mod} live{vbase} without{prep} love{nu} but{conj} they{pro} cannot{mod} earn_a_living{vp} without{prep} **hardworking{adj#vprpt_adv#_#wo/adj#vprpt}**

Because the process of tagging errors requires a reconstruction of the writer's process, the imposition of single interpretations is often not satisfactory. Although it is not our aim to indicate all possible alternatives, there is a clear need to correlate various error types and analyse recurrent patterns. The ambiguous parallel structure of the last word in the previous example requires that we assign it optional tags. Here, **adj#vprpt_adv** indicates a conjoined adjective ending in *-ing* and adverb; **#=** indicates an error in compounding; **#wo** indicates an error in word order (here an option); the symbol / indicates an alternative possibility and **adj#vprpt** indicates an adjective, with omission of verb *be(ing)*. Word components are often separated by learners when they should not be; in such cases the tag {=} is used, e.g.

Im-mature{adj#=} young{adj} couple{ns#npl} may{mod} not{adv} know{vbase} how{conj} to_spend{inf} their{padj} time{nu} in{prep} school-work{nu}

Superordinate vs. subordinate categories

At our current stage of thinking, we are using global categories based on a combination of structural and functional criteria. However, to minimise complexities we are not concerning ourselves with subdivisions such as partitives, reflexive and relative pronouns etc. Since most of these forms are unique morphological categories, they can in any event be retrieved (although of course their ellipted forms cannot).

In most other cases where use does not coincide with unique morphology, and retrieval would not otherwise be possible, we have used separate tags to distinguish, for example, existential pronouns (which appear to be overused and misused by these learners). Other areas of difficulty such as countable/uncountable nouns are also consistently noted. On the other hand, noun-verb agreement and indefinite/definite articles are tagged only in error.

Although tags are generally assigned according to the structural properties of the word-forms, we have chosen to tag both disjuncts (e.g. *obviously*) and conjuncts (e.g. *so, yet*) as {advsen}, i.e. adverbs modifying sentences. The repetitive and sometimes inappropriate use of such adverbials (cf. Milton & Tsang 1993) seems significant enough to justify a separate global category distinguishing them from other adverbs and adverb phrases.

Semantics, syntax and discourse structure

Since our tagging process is embedded in a context, it is difficult to divorce syntactic analysis from the semantic and pragmatic elements of a text or stretch of discourse. For instance, the tag *{#w}* is used when the word is inappropriate for the context, even if it falls into the correct syntactic category. Therefore, at the sentential level we are tagging within semantic and syntactic constraints, while looking beyond words as discrete units, e.g.

To_construct{inf#w} and{conj} maintain{vps} the{art} prosperity{nu} of{prep} Hong_Kong{npr} we{pro} need{vps} lots_of{quan} professionals{npl}

In this example, *to construct ... prosperity*, the collocation is unacceptable because it is syntagmatically inappropriate. The learner has apparently understood the paradigmatic correlation between *construct* and *develop*, but has not recognised the syntagmatic constraints.

In the following sentence, while there are no structural errors, the redundant article (*{art#r}*) evokes a contextually inappropriate connotation of the homonym *contact*:

Therefore{advsen} **a{art#r} contact{nc#nu}** between{prep} boys{npl} and{conj} girls{npl} is{vps} inevitable{adj}

The Learner Corpus shows the persistent problems students have with anaphoric and cataphoric reference associated with the use of determiners and pronouns. While we have not marked lexical repetition (unless incompatible with context), indeterminate co-references are tagged as deviant. Verb-tense disagreements are frequent and such contradictory shifts in tense across sentence boundaries are marked as error.

Where the structural/lexical links are so obscure that it is extremely difficult to interpret the text, the discontinuous tag *d[]d* is used. We have tried to avoid this ambiguous index as much as possible, but it is often impossible to categorise a problematic segment as anything other than a vague discursal problem, e.g.

Besides{advsen} wo[school{nadj} youngsters{npl}]wo only{adv#r} have{vps} <arti> childish{adj} outlook{nu} and{conj} childish{adj} ideas{npl},(#cs) they{pro} can't(mod-adv) encounter{vps#w} any{quan} drastic{adj} changes{npl}, **d[so{conj}] are{vps} the{art} puppy_love{nu}]d**

Although we tag formal links which disrupt unity in discourse, we do not annotate problems of coherence between sentences which are structurally and/or semantically acceptable.

Another distinctive characteristic of our interlanguage data is the abrupt and frequent shifts in register which influence lexical appropriateness. The peculiarities of spoken and written discourse overlap, as learners use items more appropriate to spoken than written discourse. Decisions regarding the status of these items are problematic. Unless meaning is obscured or syntactic errors result from register shifts, such patterns are not tagged as deviant. Similarly, phrasal/lexical redundancies are often a result of awkwardness in style and logic, and not of structural errors. These, too, are not marked as errors.

There are frequent syntactic deviations related to word-order problems. To avoid relocating words in case of word-order error, we square-bracket the segment containing the problem. Since we are not establishing constituent boundaries, it is enough to indicate the area of the word-order error by the tag *wo[]wo*.

Therefore{advsen}, they{pro} can{mod} <vbase> <expro> easy{adj} to_handle{inf#vps}
wo[people(np!) relationship(ns#np!)]wo in{prep} their{padj#artd} future{nu}

One of the major problems in tagging, and therefore accounting for error, is that it is frequently not possible to be sure of either the student's syntactic or semantic intention. In accounting for error, segments of texts are reconstructed with minimal changes to word order or shifts in syntactic categories, as far as this is possible, without obviously corrupting the semantic intention. We have tried to avoid inserting items, but rather than altering syntactic categories, we have sometimes had to insert the missing class of word or phrase in order to retain the sense of the original structures, e.g.:

In_conclusion{advsen}, two{num} more{quan} new{adj} institutions{np!} are{vps}
<arti/prep> benefit{nu} to{prep} the{art} whole{adj} society{ns}

This sentence is also an example of an error which might be tagged in several ways depending on choice of syntactic category, e.g. *institutions are a benefit to the whole society* (stylistically awkward but not strictly ungrammatical, and probably closest to the learner's construction); *institutions are beneficial to the whole society* (where the tag for *benefit* would be {nu#adj}); or *institutions benefit the whole society* (where the article and copula would be marked {#r} for redundant and *benefit* tagged as {nu#vps}). We chose the first way of 'restructuring' as it involves minimal change to the learner's text. While we try, as explained, to include obvious alternative reconstructions, it is impossible to claim that we have accounted for every possible error. More detailed alternative categories can be determined in the light of frequency data from the corpus, and depending on the number of alternatives we might want to offer students.

Omissions are inserted where they apparently ought to be, using diamond brackets (< >), e.g.

There{expro} are{vps/#agr} inside{adj} **force{ns#npl/#arti} and{conj} outside{adj} force{ns#npl/#arti} <pro>** to_make{inf#w#vps} the{art} boys{npl} and{conj} girls{npl} have{vps#r} such{pro} <arti> idea{ns}

In this example *inside force* may be a reference to one or various forces. Therefore, the alternative is offered {/}. The insertion of the second article <arti> refers back to a single idea mentioned by the learner; therefore, no alternative is offered.

Omissions beyond the word level are tagged using phrasal categories <vp>, e.g.

<vp> Any{det} romantic{adj} relationships{npl} before{prep} Form_7{npr}?

The tag marking redundancy is also used when syntactic errors result from our own reconstruction,⁶ e.g.

It{pro} can{mod} be{vbase} characterized{vpptpas} by{prep#w} quiet{adj} comfortable{adj} **and{conj#r} a{art#r} feeling{vprpt#r} of{prep#r} freshness{nu#adj}**

Our reconstruction of this sentence, *It can be characterized as quiet, comfortable and fresh*, results in several of the original words becoming redundant.

The unsettled relationship between grammaticality and acceptability has, as one would expect, generated considerable problems. We use the COBUILD dictionary as an authority on acceptability (built as it is on authentic use) whenever possible and other authoritative grammars (e.g. Quirk et al. 1985) to label syntactic categories. Inevitably, however, native intuition is used as the ultimate determiner of error.

Punctuation

Punctuation marks are not regularly tagged, because of our decision not to tag morphologically unique items,⁷ but misuses resulting in errors in contraction and possession are tagged. Such errors, together with incomplete or irregular constituents caused by sentence fragments and comma splices, comprise the main error-types relating to the use of punctuation.

Errors associated with sentence fragments ({sf}) and comma splices ({cs}) are important because they often result in obfuscation of meaning.

{cs}: Besides{advsen} wo[school{nadj} youngsters{npl}]wo only{adv#r} have{vps} <arti> childish{adj} outlook{nu} and{conj} childish{adj} ideas{npl}.{#cs} they{pro} can't{mod-adv} encounter{vps#w} any{quan} drastic{adj} changes{npl}, d[so{conj}] are{vps} the{art} puppy_love{nu}]d

{sf}: Added_to_this{advsen}, many{quan} problems{npl} may{mod} also{adv} arise{vps} because{conj} it{pro} is{vps} a{art} new{adj} institute{ns}.{#sf} Unlike{prep} other{det} older{adj} institutes{npl} such_as{conj} the{art} University_of_Hong_Kong{npr} or{conj} the_{art}Chinese_University_of_Hong_Kong{npr#cap}

There has, however, been no attempt to tag positions of omitted commas, as such omission is rarely unambiguous. By using the tags {#=} and {#_} to denote error in contraction, we are implicitly tagging for inappropriate spacing, e.g.

He{pro} is{vps} a{art} **hard_working{adv_vprpt#adj#=} man{ns}**

— where *hard working* should be *hardworking*. Irregular use of hyphens is also tagged as error, e.g. *Im-mature{adj#=}*. As mentioned, we tag expressions hyphenated in error by students as collocational errors.

The use of the apostrophe to mark the genitive inflection is tagged both in cases of omission and correct usage ({pos}). An example of omission is: *peoples{nadj#pos} views{npl}*. Negative and verb contractions are indicated ({~}), e.g. *isn't{aux~adv}* and *there's {expro~vps}*. Apostrophe omission is tagged as error.

A persistent error in question forms occurs where the student ends the sentence with a question mark but has not inverted the subject and verb. In this case, the tag *wo[]wo* is used, e.g. where the student has entitled his paper:

wo[Romantic{adj} relationships{npl}] is{vps#agr}wo better{adj} after{prep} Form_7{npr}?

What next?

We intend to proceed with compilation and analysis of the corpus. We expect to improve the speed and consistency of word-class tagging by using the University of Lancaster's automatic tagger, *CLAWS* (cf. Garside 1987; Kirk, this volume), and manually nesting error-tags within the automatically assigned word-class tags. In this way it is feasible to aim at annotating a much larger portion of the corpus than we could if we proceeded by manual tagging alone. The evidence of the last several years of corpus linguistics suggests that a corpus of anything less than 1,000,000 words is not a reliable representation of a language (or interlanguage in this case).

We would like to generate enough data about this interlanguage to adapt *CLAWS* to tag error reliably. Now that our tagged database is in the 50,000-word area suggested by Sampson (1987) for the creation of a proto-tagger, it may be possible to move on to adapt current tagging algorithms for this corpus. Atwell & Elliot (1987:122) report that *CLAWS* is already capable of diagnosing error in “ill-formed” English involving “unlikely tag co-occurrences”. We hope to be able to test the success of similar procedures in identifying the loci of error in learners’ English.

In identifying the variables that affect error and arriving at some type of classification of error gravity, it is tempting to hypothesise how the variables we have captured affect student writing. For example, we might assume that students practise avoidance strategies in the examination, and that the examination will act as a filter for fossilisation, trapping persistent error, i.e. common errors in the examination will tend to be those that are the most fossilised. However, persistent errors are of at least two types — those which students recognise as error, without being able to correct them, and those which students do not recognise as error. Presumably, students tend to avoid those constructions which they know (or are told) are ‘difficult’. It may be that the common syntactic errors in the examination are those which students (and their teachers?) have the most difficulty recognising as error. These and other speculations about the variation in error between the examination and the untimed assignments should be more accessible to discussion once we have tagged significantly large and varied corpora of writing of both types.

We do not expect it to be particularly revealing to look for patterns of syntactic or rhetorical emulation in environments beyond their school and tutorial classes, as students report reading, or listening to, English little (cf. Milton 1992). It is unlikely that we will find much of an influence on their writing from the local newspapers, for example.

Conclusion

We have reported here a procedure for tagging interlanguage and some of the problems encountered in creating a tagset. It is our intention to move on from this stage to provide an index of the syntactic patterns used by Hong Kong learners, work that we hope will have practical benefit for teachers and researchers. This study will confirm the degree to which various lists compiled from teachers’ observations, outlining some of the frequent errors of Hong Kong students, are reliable. We hope to provide information about the relative frequency of these errors — how often students get the structures right as well as how often they get them wrong.

The electronic ‘writing assistant’ that we hope to produce will provide advice at the process stage of students’ writing as well as allow students to control and structure

their proof-reading. We intend it to recognise at least the most common grammatical features a student controls and misuses or overuses. We do not imagine that machine-assisted learning will replace human teachers, nor do we expect electronic writing tutorials to be able to analyse student writing at much more than a surface level (at least not unless and until there is some major breakthrough in natural language processing techniques). Assuming that we can achieve an acceptable degree of reliability in the analysis of certain error-types, and a productive way of automating advice to students, we should be able to help students at the point they most need assistance. If we can free those who work to improve students' writing from some of the (probably wasted) drudgery of surface-feature correction — on which there is far too much emphasis anyway — our own drudgery will have been rewarded. A pedagogically effective program may take some time to develop fully, but for it to be reliable it must be based on research of the type described here.

Notes

1. Although students are required to enrol in the English course on the basis of their *overall grade* on the Use of English Examination (which comprises five sections, including the composition) the Examinations Authority reports a high correlation between the grades students receive on the composition section of the examination and their overall grades, so that we can be reasonably sure that the corpus is statistically representative of the students required to take the course.
2. Sinclair (1991b:5) advises thorough exploitation of a corpus before imposing tags on the raw data. We recognise the logic of this methodology, but the conditions of our financial sponsorship were such that we had to undertake annotation at the same time as the corpus was being transcribed.
3. We follow Leech's distinction between *corpus* and *archive*: for various reasons an archive may not be truly representative of the modelled population. In our case, the texts of the 'assignment archive' were collected opportunistically (though not always easily) and we have yet to perform the series of selection procedures necessary to carve a representative corpus from this archive. One problem we face is that much of the students' untimed, out-of-class writing is not their own.
4. Students reported in a recent survey of undergraduates (Milton 1992) that they are required to do little extended writing for their major courses, which may be a vicious circle fuelled by the perceptions of their professors that students cannot express themselves well enough to be given extended writing assignments. The latter possibility is suggested in a survey of the staff teaching undergraduate courses in the students' major areas (*ibid.*).

5. The belief by the Hong Kong Government and university educators that students are unprepared for tertiary study in English is evidenced in the enormous amount of money being spent on English 'remediation' by tertiary institutions in the territory.
6. This 'creation' of error in the reconstruction of text may seem like the imposition of double jeopardy on the error data, but the distinction between genuinely occurring error and these once-removed errors is apparent in the retrieved data, and thus can be accounted for as such in any discussion of theory or pedagogy.
7. There are, of course, few morphologically unique symbols in English. Abbreviation marks cannot be counted as separate from full stops unless tagged as such, but we have found very few cases in this corpus where such a distinction is necessary.

Acknowledgements

We gratefully acknowledge the generosity of the Hong Kong Examinations Authority in providing scripts for our corpus, and we thank our colleagues in the Language Centre, Hong Kong University of Science and Technology, for their patience in helping us to compile an archive of student assignments. We appreciate the support shown this project by the Senate Research Committee of HKUST in providing funds (Grant No. DAG92/93.LC01) to help collect and analyse the data. For the transcription of the examination scripts, much praise is due Warqa Milton.