English Profile Journal. (2011), **2**(1), Page 1 of 17, **e1**
doi:10.1017/S2041536211000018

Semi-automatic ESOL error annotation

Øistein E. Andersen Cambridge University, Cambridge, UK

Abstract

Manual error annotation of learner corpora is time-consuming and error-prone, whereas existing automatic techniques cannot reliably detect and correct all types of error. This paper shows that the two methods can successfully complement each other: automatic detection and partial correction of trivial errors relieves the human annotator from the laborious task of incessantly marking up oft-committed mistakes and enables him or her to focus on errors which cannot or cannot yet be handled mechanically, thus enabling more consistent annotation with considerably less manual time and effort expended.

Keywords: annotation, error annotation, manual, automatic, semi-automatic, learner corpora, error corpora, error detection, error correction, annotation consistency, annotation tool

1. Status quo

The number of sizeable error-annotated corpora remains limited, at least partly because 'error annotation is one of the most tedious, subjective, time-consuming and labo[u]r-intensive aspects of corpus annotation' (Wible *et al.* 2001: 311). The Cambridge Learner Corpus (CLC: see http://www.cambridge.org/elt/catalogue/subject/custom/item3646603/) is, with over $16^1/_2$ million words error-coded, one of the largest corpora of this kind currently in existence. Originally intended to inform dictionary compilation and textbook writing, it incorporates material written during language examinations by learners of English at different levels and from all over the world.

The error annotation in this corpus is added manually by a small team of trained annotators who type SGML tags and proposed corrections using a standard text editor, followed by a post-editing step (referred to as 'detoxification' by the error coders) designed to detect not only occasional SGML errors, but also inconsistencies in the annotation. Other factors contributing to increased annotation quality include the reasonably thorough coding manual and the fairly pragmatic and linguistically superficial set of error tags 'designed in such a way as to overcome[. . .]problems with the indeterminacy of some error types' (Nicholls 2003: 572). In comparison, error taxonomies used for annotating some other corpora, for instance the International Corpus of Learner English (ICLE) developed at the Université catholique de Louvain (tagset can be found in MacDonald Lightbound 2005), the Japanese

Standard Speaking Test (SST) corpus (Izumi *et al.* 2004) and the Chinese Learner English Corpus (CLEC, Gui & Yang 2002), are organised according to more linguistically motivated criteria, which, although many of the resulting error categories are the same as those found in a taxonomy based more on surface forms, tend to define certain error types in a way more dependent on the wider context and the writer's intention.

The error annotation in the CLC was originally added by one person; with the expansion of the team, this person no longer does any of the initial annotation, but instead reads through all the annotated scripts to assure a high level of consistency. This means that all texts are checked twice, first by a member of the team and then by the head annotator, who in particular looks out for inconsistencies and makes decisions as to what should be considered incorrect and marked up as such, as well as how errors should be classified and corrected when there is no single obvious solution. This approach combined with detailed guidelines should limit the amount of inconsistency caused by annotator disagreement (*cf.* Andreu Andrés *et al.* 2010, Tetreault & Chodorow 2008), but it remains difficult to obtain a quantitative measure of annotation quality in the absence of formal inter-annotator agreement studies and 'with no commonly accepted standards' for the task (Leacock *et al.* 2010: vi).

It should also be noted that certain errors are not marked up as such in the CLC, in particular misspelt proper names (apart from the ones any student of English should be able to avoid, *e.g.*, **Ingland* and **Amrica)* and errors which appear to be directly induced by the exam question. This is a sensible compromise seen from a textbook writer's perspective, but it inevitably causes certain classes of incontestable errors not to be marked up and thus to remain indistinguishable from correct text, which makes the corpus less suitable for other applications (*e.g.*, training and testing of an automatic error detection system). The practice of silently altering the text in certain cases is also suboptimal because it makes it difficult to tell exactly what was part of the original examination script and what has been modified later. An example of this is the somewhat crude anonymisation technique that consists in substituting a number of *x*s for proper names and other potentially sensitive information (given the kind of data that is typically removed by this means, simply considering *x*s as a noun phrase often works reasonably well in practice, however).

2. Consistency checking

To get an idea of the level of consistency of the error annotation in the CLC, the detection rate for simple errors was investigated: we identified words and phrases often marked up as erroneous, either unconditionally or in given linear contexts, and looked at the remaining identical occurrences to see whether those were actually wrong as well and should have been marked up. As the first few lines of Fig.1 suggest, clear errors are typically marked up as such with reassuring consistency. However, a trivial error like **occured* spelt with one *r* has actually been missed 15 per cent of the times it occurs, which seems to indicate that a system for marking up simple errors automatically could usefully complement the human annotator by spotting, in particular, typographical errors and others that are easily overlooked.

Furthermore, it turns out that certain trivial errors are inordinately frequent (*cannot* incorrectly split into **can not* alone accounts for 0.2 per cent of the errors), which implies

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

Original	Correction	Rate
accomodation	accommodation	99%
a lots of	a lot of	99%
forward to hear	forward to hearing	99%
Your faithfully	Yours faithfully	>96%
appreciate if	appreciate it if	96%
to spent	to spend	95%
center	centre	91%
However there	However, there	87%
a part time	a part-time	86%
occured	occurred	85%
On one hand	On the one hand	60%
third world	Third World	57%
other hand I	other hand, I	~50%

Figure 1 Frequently annotated errors in the CLC with correction rate, the proportion of incorrect occurrences of a word/phrase that are actually marked up. (Some of the words/phrases that should typically be corrected may be correct in specific contexts; such instances were before the rates were calculated.)

that even a relatively crude system would be able to deal with a meaningful subset of the errors and let the human annotator concentrate on more interesting/complex ones.

As for more subtle details, upon which style guides are likely to disagree, the lower consistency rates arguably indicate that the corresponding putative rules are not universally followed; for example, whether or not *Third World* should be capitalised is purely conventional. Similarly, there is no obvious reason for requiring a comma in *on the other hand, I want to improve my conditions of employment*, but not in *on the other hand I agree with the complaints*, though it is difficult to tell from the corpus whether the annotators have attempted always to require a comma after sentence-initial adverbials and have occasionally failed to do so – or whether they only require it when its absence would cause ambiguity, at least locally (as in 'garden path' sentences) and a few commas have been added which are not strictly necessary according to that approach. Another possible explanation is that the policy may have changed as a consequence of a shift in usage observed amongst professional writers and publishers (this was reportedly the case for hyphens in attributive compounds, which are no longer considered compulsory in general). Whatever the cause might be, inconsistencies make the corpus less suitable as training data or as a gold standard of errors for an automatic system to detect; such issues are to a certain extent inherent in the task of error annotation, but it is to be hoped that more sophisticated consistency checks can contribute to the detection of current inconsistencies, leading the way to clearer guidelines, at least some of which may be enforced mechanically.

3. Automatic pre-annotation

Despite considerable work on methods and systems for detection and correction of spelling and grammar errors, none of the existing error-annotated corpora seem to have been prepared using such techniques. One reason for this may be that generally available spelling and grammar checkers are made with competent educated adult native speakers in mind

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

and often unable to detect errors typical of children (Hashemi *et al.* 2003), second-language learners (Liou 1991) or even native-speaker university students (Kohut & Gorman 1995, Kies 2008). Another issue is that a tool for error annotation should not only detect an error, but also, whenever possible, classify it and provide a suitable correction.

In order to investigate the potential of semi-automatic annotation in terms of making the human annotator's task less laborious and repetitive, a system was developed that aims to detect relatively trivial errors automatically and add the corresponding annotation, including corrections when appropriate. The error detectors are largely opportunistic: no attempt is made to find all errors that could potentially be detected by a machine; instead, we focus on recurrent errors in the CLC and on those that can be identified and corrected with a high degree of confidence by exploiting information found in a dictionary. This leaves much room for improvement, but will enable us to investigate the potential of semi-automatic annotation.

3.1. Purveyors of perplexity in perpetuum (annotator 0)

Many trivial errors are committed — and corrected — over and over again, such as the ones shown in Fig.1. Our first error detector can identify many such errors by using rules derived directly from the existing error annotation: a correction rule was created for errors that appear at least 5 times and are corrected in the same way at least 90 per cent of the time. In addition to the original text marked up as erroneous, up to one word on either side was used to model the immediate context in which an error occurs. For instance, the correction *I* <SX>*thing*|*think*</SX> *that* would give rise to four potential indicators of error, *thing*, *I thing*, *thing that* and *I thing that*, each of which can be searched for and counted in the corrected text. In this case, the result of such an investigation would be that at least *I thing that* is non-existent or extremely rare in the corrected text and thus a good indicator of error, and furthermore that the error is always or most of the time corrected in the same way (*i.e.,* to *I think that*). The conclusion would be that an automatic system ought to hypothesise every occurrence of *I thing that* as a misspelling of *I think that*.

The following examples illustrate the kinds of error that can be detected and corrected using such simple rules. In particular, the previously discussed example appears as *I* <SX>*thing*|*think*</SX> *that*, which means that any occurrence of *I thing that* will result in *thing* being marked up as a spelling confusion (SX) error for *think*.

No context:
<S>accomodation|accommodation</S>
<SA>center|centre</SA>
<RP>french|French</RP>
<RP>an other|another</RP>
<UP>I'am|I am</UP>
<MP>above mentioned|above-mentioned</MP>
<W>be also|also be</W>
<ID>In the other hand|On the other hand</ID>

Left context:
the <RP>internet|Internet</RP>
reason <RT>of|for</RT>

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

all <AGN>kind|kinds</AGN>
I <SX>though|thought</SX>
despite <UT>of|</UT>
computer <RN>programme|program</RN>
to <DV>complaint|complain</DV>

Right context:
<DA>Your|Yours</DA> Sincerely
<AGD>this|these</AGD> things
<MP>long distance|long-distance</MP> travel
<UV>be|</UV> appreciate
<RV>loose|lose</RV> their

Left and right context:
50 <MP>years|years'</MP> experience
I <SX>thing|think</SX> that
I <DV>advice|advise</DV> you
a <DJ>slightly|slight</DJ> increase
is <SX>to|too</SX> small

Recurrent errors such as these can be identified automatically, but we do not want to limit an automatic annotation system to cases with 100 per cent correction rate in the CLC: first, there are imperfections in the corpus, and we should not want to discard a potential correction rule handling *I thing that* just because one occurrence of the incorrect phrase might have made it past the annotators' eyes unnoticed; secondly, a rule may be useful even if it occasionally introduces incorrect error mark-up which will have to be removed by the human annotator, as would be the case for an unconditional rule hypothesising *can not* as a misspelling of *cannot*, which would be wrong in cases like *can not only . . . , but also*. However, human annotators reportedly find spurious errors introduced automatically particularly annoying, so an imperfect rule should only be considered when the resulting annotation is correct in an overwhelming majority of the cases (incorrect instances of *can not*, in CLC outnumber correct ones by almost two orders of magnitude).

Manual evaluation of specific rules might be worthwhile if an automatic annotation system is to be employed on a large-scale annotation project, but would clearly require a fair amount of work by someone who can make policy decisions on what should and should not be marked up as erroneous, and was not feasible within the scope of this study. We instead had to apply a simple rule and chose a threshold of 90 per cent correction rate as a compromise between coverage and precision. Unfortunately, the threshold chosen precludes some obvious errors from being identified (*e.g.*, **occured*), but a lower threshold could easily lead to too many spurious errors for the human annotator to remove, and some of these errors will in any case be identified by other methods, as described in the following sections.

3.2. Morphological metamorphosis (annotator 2a)

The corpus-derived rules described in the previous section work well for specific words which are both frequent and frequently misspelt in the CLC, but do not generalise to similar or even

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

virtually identical errors involving different lexical items. Travel and tourism seem to be a popular topic in Cambridge examinations, so the misspelling of *travelled* as **traveled* with one *l* is amply exemplified, whereas **signaled* occurs only once, so no corresponding correction rule will be generated when using the proposed method and thresholds. Similarly, no rule can be derived from the corpus for a word like **groveled*, which does not occur at all, but might well appear in the future. These errors all have to do with the British English rules for *l*-doubling in morphological derivatives, and they can therefore be handled systematically, provided we have access to a word's correct morphology.

Without trying to make the corrected exam scripts conform in all respects to Cambridge University Press's house style, the CLC annotators naturally use Cambridge dictionaries to settle any doubts regarding orthography and morphology (albeit reluctantly in cases where the most recent editions do not yet reflect what is about to become established usage). It would therefore be preferable to use a Cambridge dictionary as the basis for automatic annotation rules; unfortunately, though, the ones available to us do not contain sufficient machine-readable data on inflectional morphology, so we had to use a different data source and chose the Lexical Database developed by the Dutch Centre for Lexical Information (CELEX), which in addition contains useful information on noun countability and derivational morphology.

The examples below illustrate the types of error that can be automatically detected and corrected by predicting systematic morphological anomalies modelled on actual errors found in the CLC.

Non-existent plurals (singulare tantum):
<CN>abhorrences|abhorrence</CN>
<CN>bigamies|bigamy</CN>
<CN>blamelessnesses|blamelessness</CN>

Derivation of adjective:
<DJ>academical|academic</DJ>
<DJ>atypic|atypical</DJ>
<DJ>cheerfull|cheerful</DJ>
<DJ>non-legal|illegal</DJ>
<DJ>inlegible|illegible</DJ>
<DJ>unmature|immature</DJ>
<DJ>inpossible|impossible</DJ>
<DJ>inrational|irrational</DJ>
<DJ>uncommissioned|non-commissioned</DJ>
<DJ>incertain|uncertain</DJ>

Derivation of adverb:
<DY>abnormaly|abnormally</DY>
<DY>academicly|academically</DY>
<DY>accidently|accidentally</DY>
<DY>accuratly|accurately</DY>
<DY>angryly|angrily</DY>
<DY>barily|barely<DY>
<DY>closelly|closely</DY>
<DY>wishfuly|wishfully</DY>

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

Adjective inflection:
<IJ>biger|bigger</IJ>
<IJ>brainyer|brainier</IJ>
<IJ>crazyest|craziest</IJ>
<IJ>grimest|grimmest</IJ>
<IJ>Chineses|Chinese</IJ>

Noun inflection:
<IN>addendas|addenda</IN>
<IN>addendums|addenda</IN>
<IN>alumnas|alumnæ</IN>
<IN>anthologys|anthologies</IN>
<IN>antiheros|antiheroes</IN>
<IN>bagsfuls|bagsful</IN>
<IN>boleroes|boleros</IN>
<IN>nucleuses|nuclei</IN>
<IN>oxes|oxen</IN>
<IN>schemas|schemata</IN>
<IN>tooths|teeth</IN>
<IN>aircrafts|aircraft</IN>

Verb inflection:
<IV>abandonning|abandoning</IV>
<IV>abbreviateing|abbreviating</IV>
<IV>abhoring|abhorring</IV>
<IV>abolishs|abolishes</IV>
<IV>accompanys|accompanies</IV>
<IV>amplifis|amplifies</IV>
<IV>abolishd|abolished</IV>
<IV>abolisht|abolished</IV>
<IV>abstainned|abstained</IV>
<IV>accompanyed|accompanied</IV>
<IV>ferrid|ferried</IV>
<IV>airdroped|airdropped</IV>
<IV>breeded|bred</IV>
<IV>slidden|slid</IV>

3.3. Spell-catching (annotator 2b)

The CELEX database distinguishes between British and American spellings, so a list of American words which do not exist in British English can be derived as well:

<SA>britches|breeches</SA>
<SA>jewelry|jewellery</SA>
<SA>maneuver|manœuvre</SA>
<SA>Cesarean|Cæsarean</SA>

Proper names and other words always written with a capital letter were extracted from the database to deal with capitalisation errors:

<RP>gouda|Gouda</RP>
<RP>teutonic|Teutonic</RP>

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

<RP>euclid|Euclid</RP>
<RP>scotland|Scotland</RP>
<RP>christmastime|Christmastime</RP>

Finally, a list of correct word forms was extracted from the database to enable detection of mundane spelling errors: words not already corrected in one of the previous steps and not in the wordlist can be identified as likely typographical errors. For such words, the correct spelling is unknown, and a distinguished token *(¿?)* takes the place of a correction to indicate this. It would of course be possible to make the system propose a plausible correction, for instance by using methods like the ones proposed by Deorowicz & Ciura (2005) to model the kinds of errors typically committed, relying on statistics from the CLC for error frequencies. This would clearly be useful in a tool aimed at less confident language users and would be an interesting extension to the system, but seems less important in the context of error annotation and is unlikely to have a significant impact on the annotation speed given that all frequent errors with obvious corrections will have been handled by the corpus-derived rules. Words containing at least one capital letter are not considered here, partly because we have not tried to compile a comprehensive lexicon of names, partly because of the CLC policy of generally not correcting proper names.

3.4. Euphonia (annotator 3)

The correct choice between the two euphonic variants of the indefinite article *(a/an)* depends on the following sound, the well-known rule prescribing *a* before a consonant and *an* before a vowel. The CELEX database provides pronunciations and we were already using the database for other error types, so it seemed natural to take advantage of that information. Only the first sound in a word is significant for the choice of *a* or *an*, and only whether it is consonantal or vocalic, as illustrated in the following examples:

minister consonant
MP vowel
open vowel
one consonant
home consonant
hour vowel
hotel vowel/consonant
utter vowel
useful consonant
Uruguay vowel/consonant

Words with alternative pronunciations, such as *Uruguay*, may be used with either form of the article. (The traditional usage of *an* in front of unaccented aspirated *h* is not currently accounted for, but the information needed is available in the CELEX database, so this could easily be added.) The text is part-of-speech-tagged with RASP (Briscoe *et al.* 2006) before this step to distinguish between the definite article and other instances of *a/an* (such as the *A* s in *an A in Drama is apparently as valuable as an A in Greek*).

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

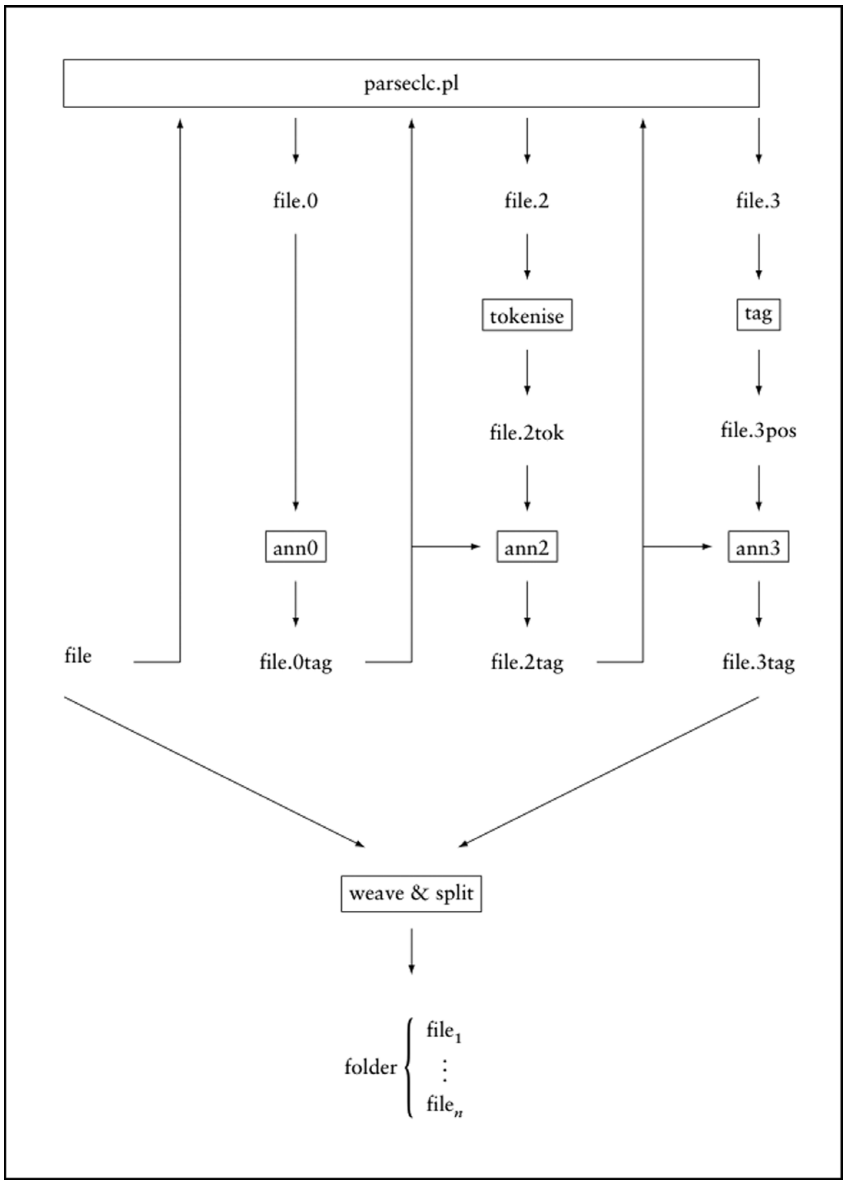

Figure 2 Schematic overview of the automatic annotation process, starting with a single file containing multiple unannotated exam scripts and ending up with a set of files, each containing an annotated script.

3.5. Synopsis

The flowchart in Fig. 2 illustrates how the different parts of the system interact to produce automatic annotation.

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

Each exam script in the CLC contains information about the candidate and the exam taken, as well as the actual text written:

```
<candidate>[metadata about examination and examinee]</candidate>
<text>
  <answer1>
    <question number>9</question number> [...]
    <original answer>
      <p>I wont be going to the little nice persian caf&eacute;
        this after noon becauce I eated to much for lunch
        and have now a awfull stomachache.</p>
    </original answer>
  </answer1>
  [more answers]
</text>
```

A tool called *parseclc* extracts the text to analyse:

```
<p>I wont be going to the little nice persian café this after noon becauce I eated to much for lunch and
have now a awfull stomachache.</p>
```

The only detail worth mentioning at this point is the normalisation of *é* to *é;* XML provides many ways to represent a given character, and ensuring that *é* never appears as for instance *é* or *é* simplifies further processing. The first set of error tags are then added by the first mechanical annotator, *annotator 0*, which uses simple string matching to detect frequently recurring errors:

```
<p>I
<e t = "MP"><i>wont</i><c>won't</c></e>
be going to the
<e t = "W"><i>little nice</i><c>nice little</c></e>
persian café this
<e t = "RP"><i>after noon</i><c>afternoon</c></e>
<e t = "S"><i>becauce</i><c>because</c></e>
I eated <e t = "SX"><i>to</i><c>too</c></e>
much for lunch and
<e t = "W"><i>have now</i><c>now have</c></e>
a
<e t = "DJ"><i>awfull</i><c>awful</c></e>
<e t = "RP"><i>stomachache</i><c>stomach ache</c></e>.
</p>
```

At this point, error tags have been added, which will have to be removed before the next processing step. *parseclc* again extracts the text, using the corrections rather than the original text when applicable:

```
<p>I won't be going to the nice little persian café this afternoon because I eated too much for lunch and
now have a awful stomach ache.</p>
```

This partly corrected version of the text is then passed through RASP's sentence splitter and tokeniser, providing input to the second annotator, *annotator 2*, which detects morphological

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

and typographical errors. The process is repeated once more, this time adding part-of-speech tags for *annotator 3* to be able to detect article form errors, which gives the following annotated output:

```
<p>I <e t = "MP"><i>wont</i><c>won't</c></e>
be going to the
<e t = "W"><i>little nice</i><c>nice little</c></e>
<e t = "RP"><i>persian</i><c>Persian</c></e> café this
<e t = "RP"><i>after noon</i><c>afternoon</c></e>
<e t = "S"><i>becauce</i><c>because</c></e>
I
<e t = "IV"><i>eated</i><c>ate</c></e>
<e t = "SX"><i>to</i><c>too</c></e>
much for lunch and <e t = "W"><i>have now</i><c>now have</c></e>
<e t = "FD"><i>a</i><c>an</c></e> <e t = "DJ"><i>awfull</i><c>awful</c></e>
<e t = "RP"><i>stomachache</i><c>stomach ache</c></e>.
</p>
```

Because the XML mark-up is handled properly, the fact that *awful* is embedded within an error tag does not prevent the system from detecting that the preceding determiner should be *an* rather than *a*. If *parseclc* had been applied again, the following output would have been generated:

```
<p>I won't be going to the nice little Persian café this afternoon because I ate too much for lunch and
now have an awful stomach ache.</p>
```

The process can obviously continue with, for instance, the generation of syntactic annotation as input to a subsequent automatic annotator. For the purposes of this experiment, though, the output from annotator 3 was combined with the original file (containing metadata irrelevant for the automatic error annotation) to create complete automatically annotated files for the human annotator to work on.

4. Annotation tool

Whereas some corpora have been annotated using dedicated tools such as the Université Catholique de Louvain Error Editor (UCLEE, *cf.* Dagneaux *et al.* 1998), the CLC annotators have written SGML tags directly in a text editor. This is not necessarily an impediment to efficient annotation compared to more visual systems which may require error tags to be selected from menus and submenus, for the coding scheme uses short codes and makes judicious use of SGML abbreviation techniques in order to limit the number of characters, and thus keystrokes, needed to mark up an error. The code is also quite readable as long as there are not too many nested errors, but occasional SGML errors, which render the entire file in which they occur unparseable until the error has been corrected, are nevertheless difficult to avoid completely. An additional consideration for semi-automatic annotation is the ease with which an incorrect error tag added by the machine can be removed by the human.

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

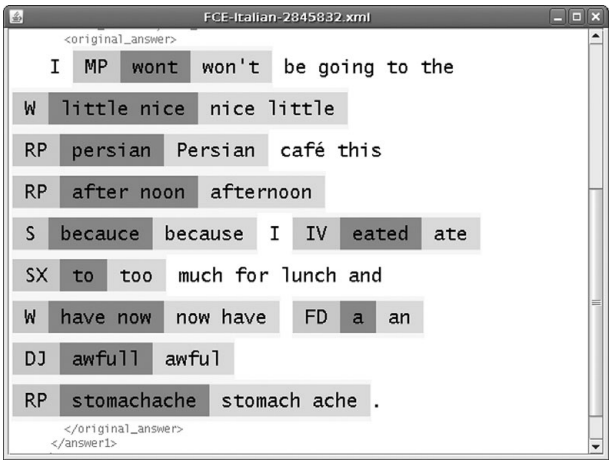

Figure 3 The pre-annotated example sentence as it appears in the annotation tool. For each error annotation, the error type is shown to the left, on an orange background; the error in the middle, on a red background; and the correction to the right, on a green background.

We felt that a simple annotation tool was the right solution: it would provide a graphical representation of the error annotation, making it easier for the annotator to see where each error begins and ends, in particular in the case of nested errors; the number of keystrokes needed could be reduced further, and the need for typing 'exotic' characters eliminated; SGML errors would never appear; and one keystroke would be sufficient to remove an unwanted error tag. Fig. 3 shows how a sentence with error annotations appears in the annotation tool.

5. Annotation experiment

The head annotator kindly agreed to try out the system outlined on the preceding pages to annotate text taken from previously unannotated parts of the CLC. We selected, for each part of the experiment, scripts from all major Cambridge examinations, which cover levels A2–C2 in the Common European Framework of Reference for Languages (CEFR). After initial testing and development, four different set-ups were tried, as described in the following sections, in order to investigate the contribution of different factors.

5.1. Manual annotation (part 1)

Statistics from previous years of CLC annotation enable us to estimate average annotation speed in terms of tags per hour or words per hour. We were concerned that those data points might not be directly comparable with the ones obtained as part of the experiment, though, and therefore included a batch of scripts for manual annotation, asking the annotator to type tags in a text editor as previously.

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

As expected, a few SGML errors appeared:

Mismatched tags:
<#DK>competitable|competitive</#DJ>
<#RJ>fashion|fashionable</#DJ>
<#SA>humor|humour</#SX>
<#FV>making|to make</#RV>

Missing angle bracket:
<#DJ>successfull|successful</#DJ

More complex error:
<#UV>I'm</#I> (should have been <#UV>I'm|I</#UV>)

There were also some overlooked errors which the automatic system would have detected (the annotator later told us that the first two errors were deliberately ignored):

<RP>clare|Clare</RP>
<RP>10'000|10,000</RP>
<SA>analyze|analyse</SA>
<SA>analyzed|analysed</SA>
an <MP>all time|all-time</MP> low
our <MP>day to day|day-to-day</MP> life
it is <SX>to|too</SX> complicated
had to <RV>seat|sit</RV> in the back row

As for the annotation speed in this experiment compared to previous annotation of the CLC, the two turned out to be significantly different (see Table 1); this can at least in part be ascribed to better English with fewer errors in the experiment (on average 1 error tag added per 14 words) than in previously annotated parts of the corpus (1 tag per 9 words) and is thus not entirely surprising, but it also shows that any direct comparison with previous years' results is likely to be misleading.

5.2. Semi-automatic annotation (part 2)

For the second part of the experiment, scripts were pre-annotated automatically using the system described in Section 3 before it was given to the human annotator for correction and supplementation using the annotation tool. Examination of the final annotation showed that the automatic pre-annotation system had precision of 93 per cent and recall of 24 per cent (see Table 2). This is quite encouraging given that the system is neither comprehensive nor fine-tuned: increased recall without loss in precision can be obtained by extending the system's coverage, and precision is impeded by an incomplete and slightly outdated lexicon. Performance is similar across CEFR levels; whether or not the system is effective appears to depend more on individual characteristics of the particular script than on the general level of English. The annotation speed turned out to be about 50 per cent higher than in the previous experiment; in addition to this, there are no SGML errors to correct, and the annotation

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

Table 1 *Performance in terms of annotation speed.*

The first two lines in the table relate to the part of the CLC that has been error-coded during the last couple of years, the figures on the first line only including the time spent on the initial coding, the second line including the subsequent post-editing step ('detoxification' to remove SGML errors and coding inconsistencies) as well; the remaining lines relate to the annotation produced as part of the annotation experiment described in this paper. The number of words and tags is indicated for each part of the corpus, and the inverse tag density (words per tag) is calculated to give an idea of the amount of errors (more words per tag means fewer errors, higher-quality text and less work for the annotator). The number of hours spent to annotate (including post-editing in the case of the second line) each part is indicated, which, in combination with word and tag counts mentioned previously, allow the annotation speed to be calculated in terms of words per hour as well as tags per hour.

	Words	Tags	Words/tag	Hours	Words/hour	Tags/hour
CLC coding	6,736,452	746,252	9	5,156	1,306	145
—— & detox	——	——	——	6,924	972	108
Part 1	13,127	934	14	4	3,281	233
Part 2	19,716	1,433	13.8	?4	4,929	358
Part 3A	9,881	311	31.7	1.51	6,544	206
Part 3B	9,679	1,023	9.46	2.71	3,572	377
Part 3 (Σ)	19,560	1,355	14.65	4.22	4,635	316
Part 4	18,610	1,373	13.55	1.66	11,210	827

is more consistent, which eliminates the need for subsequent SGML verification and vastly reduces the need for consistency checking, thus making the effective speed increase closer to 100 per cent.

5.3. Annotation of individual sentences classified as good/bad (part 3)

In order to get a better idea of how important context is for correct annotation, as well as to assess the potential for more efficient annotation by focusing on sentences more prone to contain errors, sentences were split into two sets, likely to be correct (3A) and likely to contain errors (3B), for the third part of the experiment. As one would expect, this set-up caused the annotation speed in terms of words per hour to increase for the largely correct sentences, and in terms of error tags per hour for the largely incorrect sentences, whilst both performance measures declined globally, at least partly because it is more burdensome and mentally exhausting for the annotator to deal with individual sentences than connected passages of discourse. Sentence-level performance is shown in Table 3.

5.4. Re-evaluation in context (part 4)

Finally, the sentences from part 3 were put together again and presented to the annotator anew for evaluation in context. This gave a precision figure for manual detection of errors

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

Table 2 *Performance of the pre-annotation system in terms of precision and recall measured against a human annotator.*

Precision is defined as the proportion of actual errors (identified by the human annotator) amongst the purported instances signalled by the system, while recall is the proportion of real errors (found by the human annotator) actually detected by the system. The before *column indicates the number of tags added during the pre-annotation step; the* after *column indicates the total number of errors after annotation; and the* correct *column indicates the intersection between the two sets (*i.e., *the number of tags added during pre-annotation that were not subsequently removed during annotation). Note that Part 4 uses human pre-annotation (resulting from the Part 3 annotation).*

	Before	Correct	After	*P*	*R*
Part 2	372	345	1,448	93%	24%
Part 3A	0		280		
Part 3B	397	353	1,023		
Part 3 (Σ)				89%	27%
Part 4	1,302	1,293	1,373	99%	94%

Table 3 *Performance in terms of the system's ability to detect sentences containing at least one error.*

The total *column shows the total number of sentences. The* before, correct *and* after *columns have the same meaning as in Table 2, but they refer to a number of sentences rather than a number of individual errors.*

		Tagged sentences					Classification	
	Total	Before	Correct	After	P	R	P	R
Part 3A	686	0		198		(71%)		
Part 3B	517	297	271	417	91%		81%	
Part 3 (Σ)						44%		68%

in individual sentences out of context of well over 99 per cent, whereas recall was a bit lower at 94 per cent. We can conclude from this that what appear to be errors when a sentence is regarded in isolation usually turn out to be errors in context as well, whereas about 1 out of 20 errors require extra-sentential information to be detected. No context-dependent errors were identified at the lowest level (CEFR A2), which may be related to the limited length and template-like nature of the writing, but it should also be kept in mind that the amount of data is small. The higher levels (B1–C2) all exhibited instances of context-dependent errors in this experiment, including a relatively larger proportion of verb tense errors compared to the set of errors identifiable in a sentence context, but also quite a large variety of others such as word choice and article errors.

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press

This part of the experiment also permitted us to calculate an upper bound for annotation speed given very high-quality pre-annotation: compared to part 1, there was an increase of 250 per cent in annotation speed, or towards 325 per cent if the amount of post-editing can be reduced. However, the fact that the annotator had already seen the sentences, albeit out of order, may also have contributed to the speed increase observed in this part of the experiment.

6. Conclusion

We have seen that an annotation tool that incorporates automatic techniques for error detection and correction can contribute to higher accuracy and increased productivity in the task of error annotation. This is significant since manual error annotation can be both laborious and tedious, whereas the existence of sizeable error-annotated corpora is crucial both for the study of language containing errors (be it from a pedagogic or a more purely linguistic perspective) and for the development of 'grammar checkers' and other tools that actually address the areas of language that can be shown to be problematic.

Our goal was to demonstrate that semi-automatic annotation can be beneficial, not to develop a comprehensive set of high-quality error annotators. Further work on this should permit better automatic annotation.

Finally, a suitably adapted version of the automatic error detection and correction system presented in this article can be used on its own for other applications, for instance as a tool for learners to avoid frequent errors or as a starting point for research into orthographic features of texts written by learners from different backgrounds (see Cook & Bassetti 2005).

Acknowledgements

We should like to thank Diane Nicholls for her keen involvement in this project and for all her valuable comments during the development of the annotation system, and the anonymous reviewers for their helpful remarks and suggestions, as well as Cambridge University Press for having granted access to the CLC, and Cambridge ESOL for financial support. This article reports on research supported by the University of Cambridge ESOL Examinations.

References

Andreu Andrés, M. A., Guardiola, A. A., Matarredona, M. B., MacDonald, P., Fleta, B. M. & Pérez Sabater, C. (2010). 'Analysing EFL learner output in the MiLC Project: An error *it's, but which tag?' In Campoy-Cubillo, M. C., Bellés-Fortuño, B. & Gea-Valor, M. L., *Corpus-based approaches to English language teaching*. London: Continuum, 167–179.

Briscoe, E. J., Carroll, J. & Watson, R. (2006). 'The second release of the RASP system'. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney, Australia, 77–80.

Cook, V. & Bassetti, B. (2005). *Second Language Writing Systems*, volume 11 of *Second Language Acquisition*. Multilingual Matters, Clevedon.

Dagneaux, E., Denness, S. & Granger, S. (1998). 'Computer-aided error analysis'. *System: The International Journal of Educational Technology and Language Learning Systems. 26.2*, 163–174.

Deorowicz, S. & Ciura, M. G. (2005). 'Correcting spelling errors by modelling their causes'. *International Journal of Applied Mathematics and Computer Science. 15.2*, 275–285.

桂诗春 [Gui, S.] & 杨惠中 [Yang, H.] (2002). 中国学习者英语语料库 *[Chinese Learner English Corpus]*. 上海外语教育出版社 [Shanghai Foreign Language Education Press], Shanghai, China.

Hashemi, S. S., Cooper, R. & Andersson, R. (2003). 'Positive grammar checking: A finite state approach'. In Gelbukh, A. F. (ed.), *Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003, Mexico City, Mexico, February 16–22, 2003, Proceedings*. 635–646.

Izumu, E., Uchimoto, K. & Isahara, H. (2004). 'SST speech corpus of Japanese learners' English and automatic detection of learners' errors', *ICAME Journal, 28*, 31–48.

Kies, D. (2008). 'Evaluating grammar checkers: A comparative ten-year study'. Presented at the 6th International Conference on Education and Information Systems, Technologies and Applications (EISTA), Orlando, United States. <http://papyr.com/hypertextbooks/grammar/gramchek.htm>.

Kohut, G. F. & Gorman, K. J. (1995). 'The effectiveness of leading grammar/style software packages in analysing business students' writing'. *Journal of Business and Technical Communication, 9.3*, 341–361.

Leacock, C., Chodorow, M., Gamon, M. & Tetreault, J. (2010). *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool: San Rafael, California, USA.

Liou, H.-C. (1991). 'Development of an English grammar checker: A progress report'. *CALICO Journal, 9.1*, 57–70.

MacDonald Lightbound, P. (2005). *An analysis of interlanguage errors in synchronous/asynchronous intercultural communication exchanges*. Ph.D. thesis, Universitat de Valéncia.

Nicholls, D. (2003). 'The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT'. In Archer, D., Rayson, P., Wilson, A. & McEnery, T. (eds), *Proceedings of the Corpus Linguistics conference, volume 16 of University Centre For Computer Corpus Research on Lanugage, Lancaster University, Technical Papers*. Lancaster, 572–581.

Tetreault, J. R. & Chodorow, M. (2008). 'The ups and downs of preposition error detection in ESL writing'. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, 865–872.

Wible, D., Kuo, Ch.-H., Yi, Ch. F., Liu, A. & Tsao, N.-L. (2001). 'A Web-based EFL writing environment: Integrating information for learners, teachers and researchers'. *Computers & Education, 37.3–4*, 297–315.

https://doi.org/10.1017/S2041536211000018 Published online by Cambridge University Press