USING COMPUTERS IN CORRECTING WRITTEN WORK

by James Thomas
Faculty of Informatics
Masaryk University
Brno, Czech Republic

thomas@fi.muni.cz

Introduction

Computers assist and automate many aspects of our lives and enable things that are barely possible, if at all, without them. Many forms of writing are undertaken on computers since word processing programmes allow editing and formatting, typing shortcuts, spelling and grammar checking, storing different versions, adding pictures, footnotes, hyperlinks within the same document and to the internet, as well as working online with someone at a remote computer. Collaborative writing also takes place via email, chat and computer-mediated-communication facilities such as Nicenet and the Tandem project.

When a text is submitted for correction, whether for pedagogical or proofreading purposes, it may seem anachronistic for the author to print it out for the corrections to be made manually, especially given that the same corrections and the same comments are repeatedly made and the same reference resources recommended, often themselves hyperlinks (URLs).

On another front, there has also been a considerable amount of research and development in computational linguistics, and this has led to advances in human-machine communication, translation and speech recognition software, and new empirical findings about language itself, grammar, vocabulary and their interrelationships in particular. With the advent of desktop computers, the backbone of this resource, the corpus, can be consulted by anyone correcting written work when they come across language that seems to deviate from normal usage. In fact, students involved in data-driven learning, as the pedagogical application of computational linguistics is known, make these observations themselves.

Whenever I give my teacher training seminar, Using Computers in Correcting Written Work, the participants are often disappointed to find that technology is not going to automatically resolve the linguistic questions that linger over their students' language. This article follows on from Krajka (2002), in which he describes some of the useful tools available for correcting student work with the computer. I will refer to some of these tools, some linguistic resources that both teachers and students can use today, and how in the future these could be combined into powerful intelligent tools.

Some currently available resources

Word processors

Using a word processor to correct written work has a number of advantages. Microsoft Word (henceforth MSW), the word processor described in Krajka (2002), offers spelling and grammar checking both of which offer suggestions which can be accepted or rejected. The program is intelligent enough to recognize passives, *their are* is automatically corrected to *there are*, for example. The program offers a comma before *which* although a quick look at a concordancing program will show that this requirement is overstated. Back in MSW, the user can set the stylistic level, e.g., *casual, technical*, each of which makes different demands such as the use of contractions, and the very real problem of too many successive nouns. Such intelligence can be quite useful in bringing possible errors to a marker's attention. But a comprehensive list of errors in student writing is much longer than those offered by this grammar checker. As we shall see, such a list is not the best path to identifying deviations.

Another area of intelligence that some wordprocessors offer is basic text "Statistics and Readability" scores. This does not assist you in marking, but it evidences a computer's ability to perform basic statistical functions on a text as is pursued throughout this article. The Flesch Reading Ease (FRE) is a calculation based on the average sentence length and the average word length in a text. The result is between 0 and 100, and the higher the score, the more readable the text is said to be. The Flesch-Kincaid Grade Level converts the number to an expression of U.S. grade-school level. At the Juicy Studio website, you can determine the FRE of a webpage by entering its URL. The statistics are explained in greater detail at Plain Language. Click here to see a sample of these statistics in MSW.

For marking purposes, "Track Changes" in MSW is often used. When "Track Changes" is turned on, the alternative wording which you type over text that is nominated for editing, appears beside the original which becomes crossed out. When texts are returned to their authors, both the original form and the suggestions can be seen. They can then choose whether or not to accept them. Students could blithely accept everything recommended without making any effort to understand the nature of the error or account for the correction. This is not sound pedagogical practice and "Track Changes" is more suited to proofreading.

Markin

A program called Markin (by Martin Holmes) has been created specifically for the purpose of correcting students' written work. Unlike Track Changes, this program does not encourage the marker to provide better alternatives, although they can be added as comments. Rather, when a

teacher locates an error, clicking on an appropriate button (click here to see the default button set) will automatically insert a comment in the student's text, such as Missing Word, Word Order, Article. An example of the output-as-webpage can be seen here. This requires that students discover their own solutions to the problems identified, edit their work and resubmit it. The pedagogical impact of this mode of returning written work to students is described by Chappelle (1998) thus:

> When errors are recognized in comprehensible output, the process of the learner's self-correction is also believed to be beneficial particularly because the linguistic items for which self-correction occurs may be those for which learners' knowledge is fragile. … Corrections can come from learners' own hypothesis testing, from their requests for assistance from others, or from explicit correction.

While on the topic of returning Markin texts to students, there are some practical issues to be taken into consideration. Markin does not provide the marker with any way of seeing what changes the student has made: you simply have to read the edited text as if for the first time. A split screen that allowed several editions would be very helpful, especially in the context of process writing.

Daudin-Vigot makes the following comment to her students in Making the Most of Markin Corrections that "…the html version is interactive but not editable. This is why I'm also sending the rtf version for you to make the necessary corrections." Finally, to avoid confusion when students do submit written work, instructions also need to be given regarding consistent labelling of the work.

As can be seen at the end of the output the student receives, Markin creates a table summarising positive and negative comments. It is currently possible to get a batch of statistics on a group of texts, which is a useful diagnostic tool for teaching. The group, however, does not have to consist of different students – it can be the same student on different tasks which allows an individual's improvement to be statistically depicted. I suspect that a split-screen version of the program could make even more significant statements about improvement, especially on a single process-writing task.

Markin allows the user to create new buttons for automating positive and negative feedback and the set provided contains useful standard language errors from an analytical point of view. However it does not offer any holistic marking criteria such as Reader satisfaction, Quality of argument, Style appropriateness, as belong in "end comments". Such feedback is important for

student development. Since software design influences how people interact with it, the adoption of a more sophisticated concept of marking could advance the way marking is undertaken. Global statistics and readability scores of a text may also be indicative of the overall sophistication of the writing.

MS Word and Markin offer facilities that are useful to either correcting or proofreading but not enough for dealing with both. The next section describes a program that potentially meets both requirements, but it too is not without its limitations.

Wincorr

A small program called WinCorr (Kukačka and Chalupský) has been created at the FI MU for the purpose of creating a corpus of errors that native speakers of Czech make in their academic writing. It is not currently designed to meet a wide range of potential users and consequently does not offer a wide range of options. In fact, it only exists in a Czech form at the moment. Its existence, however, invites comparison in some respects.

Once a text is loaded into the edit frame, you choose to correct an error by selecting a word or phrase. You are then offered a general menu of error types: spelling and typing, morphology or syntax, semantic or lexical, stylistic or structural, unclassified. Click here to see some screen shots.

Clicking on the relevant one of these presents you with a submenu. For example, Stylistic errors include ambiguity, errors in reference and co-reference, repeated words. The long labels make it easier to use than Markin's three letter buttons. And the linguistically-based categories draw attention to more general language areas that need attention. At the bottom of the dialogue box, you can choose how the correction should be inserted, and you can include the correct form. When printing, you can choose between including the correction data or printing only the corrected form. This combines some of the advantages of MS Word's Track Changes and Markin's error indications.

However, Wincorr's printout with the correction data is not easy on the eye, and as with Markin, the program does not exploit any linguistic intelligence.

The remainder of this article concerns language per se and how computers are employed to deal with some aspects of it. It begins with a pedagogical classification of language that derives from the lexical approach (Lewis 1993) and then exemplifies public access to linguistic resources that can be used when correcting written work.

Three categories of language error

The following sentence appeared in a recently received email from one of my ex-students:

I am not very optimistic about teaching writing any more because I find it impossible to correct competently. ... one needs to be a native speaker or very good.

Like many such teachers, she is very good, and yet not confident in this area. In some cases, such doubts are well-founded, as even native speakers can be inclined to let things through that are not quite right, often because the focus is on meaning rather than form: if the meaning is clear, it is *enough good. There is also the pedagogical consideration that too much "red pen" can be demotivating.

Patterns, Facts and Choices

Things that are indisputably wrong do not present this dilemma. In the following attested examples, the first two are errors in the application of a grammar pattern, i.e., something that holds across an aspect of language, such as regular plurals, comparative adjectives and verbs, SVO word order.

> Then the interpretation of such *situation is as follows. (missing article)
>
> … should be hyphenated as 're-cord' when it is a noun, but 're-cord' *being it a verb. (unwilling to use "when it" twice).

The next two are errors of **fact**, i.e., information about a word or structure that pertains to a single item, such as *criteria* being the plural of *criterion*, or the requirement for *put* to have an object (…) and an adverbial (destination).

> … which permits *to store any objects (permit + obj+ inf with to)
>
> The second way of* object lookup is to use the knowledge of … (way of + -ing)

Errors of pattern and fact do not present major problems for correction and should be within the realm of a computer program. The real difficulty however lies with making the best **choice** from acceptable alternatives – this ushers in the possible-probable dichotomy. John Sinclair (1991) contrasts the open-slot principle (slot and filler model) with the idiom principle. In the former, any word or phrase which can be a subject, for example, can go into that syntactic slot: the result should be a "good" sentence, as we know from Chomsky's 1957 sentence, "Colorless green ideas sleep furiously". See an article in Dictionaries for more on this. On the other hand, the principle of idiom is that a language user has available to him or her a large number of semi-

precontracted phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991:110)

The following extracts from students work exemplify this.

> … has a *really sophisticated mechanism … (inappropriate style)
>
> The strategy determines the beginning element of the … (*beginning* is understandable but *first* or *initial* is more probable)
>
> Last but not least, the final force is computed … (inappropriate style – MS Word offers Finally yet importantly)

While many successful uses of language contain improbable word combinations, particularly in creative writing, smooth reading relies on a minimum of conspicuous surprises. This leads us now to the question of whether a computer tell us which word combinations are most probable.

Distinguishing between possible and probable choices

In this section, we will look at how counting co-occurrences demonstrates likelihood and then we shall examine what is publicly available for this purpose.

This table of the frequency of co-occurrences of these very common items derives from the full British National Corpus (BNC). It demonstrates, perhaps counter-intuitively, that all combinations do occur, but some are significantly more frequent than others.

	morning	**afternoon**	**evening**	**night**
last	11	21	64	8475
yesterday	345	196	90	4
this	4082	1703	1083	85
tomorrow	411	100	70	447
in the	3691	971	1197	585
next	1234	31	57	86
at	22	2	20	3034

The full BNC includes spoken English, which is one reason *yesterday night*, for example appears at all. And *next afternoon* is almost always preceded by *the.* The other reason is that some of these pairs of words (bigrams) are not complete adverbials but parts of phrases, e.g. *at morning service*, in

> The Rev Bob Morgan said prayers at morning service at the Church of the Resurrection in
>
> Ely, just 400 yards from where Mr Reed died.

The following attested fragment contains a number of errors.

The second way of object lookup is to use the knowledge of …

The error here that demonstrates likelihood is *way of.* This two-word phrase occurs 9,677 times in the BNC. Is it possible for *way of* to be followed by a noun (group)? Yes, 1,974 (20%) of the words which follow it are nouns. And 1,013 of those (51%) are the word *life*, 30 *things*, 12 *business* and most of those remaining are gerunds, *thinking, living, understanding* being the most frequent. In fact, with *way of life* being a chunk, the 9,677 could be reduced by 1,013 to 8,664 when considering the probable colligation patterns. With only a one-in-five chance that *way of* would be followed by a noun, a more probable alternative should be considered. There are 6,027 concordances with a present participle following *way of* which is 70% (not counting *way of life*). There are even 117 concordances containing the whole frame *way of doing XX is to*. So the most **probable** way of expressing this is *The other way of looking up an object is to use …* This might not satisfy the authors if "object lookup" is a term. This would then argue for more data or using a corpus specific to the domain.

In practice, it is not necessary to calculate statistics in such detail to solve most quandaries. The point is that such observations do offer the most probable way of expressing something. Importantly here, these conclusions can be reached through publicly available web-based resources.

Web-based Resources

Dictionaries and thesauri, collocation lists and word association data are all available electronically. They are faster to use and often contain more data than print resources. However, as we shall see in some of the following examples, they cannot resolve all conundrums. Since the most recently published dictionaries and grammars are distillations of corpus data, it is helpful to directly consult the data directly.

This information is available to everyone via the Internet. In this section, I will refer to two sites, namely, **Bonito** and **Just the Word**, both of which use the BNC as their corpus. To access the BNC through Bonito, it is necessary to fill in an online registration form.

Bonito

Bonito (Pavel Rychly) shows us, for example, that *needed* in *allow all needed manipulations*, an attested example, is improbable. A search of *needed* as an *adjective* returns 33 concordances, of which only eight are in fact adjectives. Entering *all needed* into Word Phrase, returns nine concordances, and all are verbs.

In the same article, the student wrote …*is very time and cost consuming*. Entering *cost consuming* into Word Phrase returns this improbable collocation once, but it is clearly spoken language and the speaker is experimenting with alternatives, as can be seen in the screen shot here.

When a student wrote *with limits given by*, the improbable *given* needed to be replaced. By entering *limits* into the lemma space and choosing *noun*, it found 6,300 concordances. To find the most probable verb to follow it, click on Freqs, and enter 1R (first word to the right) and choose Lemma. The first verb in the resulting list after *be* is *set* while *bring* does not occur once.

Just the Word

Just the Word (Pete Whitelock) is another online facility that searches the BNC and provides statistically based lists of co-occurring items, both grammatical (colligation) and lexical (collocation). In the process of proofreading an academic article, I came across *primary advantage*. While the meaning is abundantly clear, it did not feel quite right. A quick search in Just the Word did not find it. When you search Just the Word for *advantage*, the significant ADJ N pairs are presented in nine semantic clusters. Click on the relevant hyperlinked colligations in the right pane to go to the clustered collocations. Click on any of those collocations in the left pane to see the actual concordances. Try it!

Both of these programs are very flexible, allowing an infinite number of searches. In the context of non-native speaker writing, they can be of much assistance in finding the most probable word choice and colligation. From just a few examples, we can see that these user-friendly resources can help us resolve the probability of a construction in question. Students could of course turn to these resources themselves in the process of writing, and much is written these days about this discovery approach, namely Data Driven Learning (DDL). For a portal into DDL, see Tim Johns,

the father of DDL, and the DDL page within Gateway to Corpus Linguistics on the Internet. For a discussion of the approach, see Data-Driven Learning (DDL): the idea, by Bernd Rüschoff.

Before launching into the future, there is one other existing online resource that I would like to refer to. This is in the context of returning written work to students in a form that promotes a discovery approach.

Returning work to students

When students receive their texts back corrected using Markin, the location and nature of errors have been indicated, but alternatives have not usually been suggested. One way students can explore their errors is to look at their text through the window of a corpus, i.e., to see concordances of specific words. The online resource, Compleat Lexical Tutor (Tom Cobb, University of Quebec), has a facility for creating a hyptertext version of a text. This means that when users double click on any word in the hypertext, concordances exemplifying how that word form is used in the Brown Corpus are displayed in the lower frame of the screen. From this, the solution to the language problem can often be found. The right pane displays the WordNet dictionary entry.

For the teacher, creating the hypertext involves clicking on Hypertext Builder, pasting in the student's text and naming it. Click Build and its name appears which you then click to visit the hypertext page. Provide the student with its URL by pasting it into the Markin version of the student's text. It literally takes seconds to do all of this.

The concordances that the Hypertext Builder present have some limitations: they derive from the Brown Corpus which is small and old in comparison with modern-day corpora, and secondly, the search word is not lemmatised. This means that it does not find *replaying, playfully, exams* and *dictations* as wordforms, even though *play, exam* and *dictation* are in the corpus. In this situation, the user can perform a manual search on a base wordform.

What the future holds

What clearly emerges from my in-service seminars is that teachers would appreciate a program with some linguistic intelligence. Markin, Wincorr and the grammar checker in MSW are all useful tools, and combining their advantages in one program would be very beneficial. But the addition of linguistic intelligence would be a great boon to the field. It exists in speech recognition programs, for example, Dragon (from Scansoft): speak to your computer and the words appear on the screen – it accurately distinguishes between *their* and *there,* for example,

through grammatical analyses and awareness of context. But none of these programs indicate that any preposition after *accustomed* apart from *to*, or that *way how*, in *the way how to do something* are perilously unlikely.

As we saw in the previous section, corpora will provide us with data, but only if we ask for it: we then have to interpret it. It would be preferable if a program could indicate all the instances of unlikely word usage in a document, by identifying the type of document and then by comparing word combinations that are typical for that type.

In reference to the table of time words above, if a students writes, … *and next afternoon I will go football*, an intelligent program could tell us that *next afternoon* is mostly preceded by *the*, but when followed by a future clause, is probably *tomorrow afternoon*. Furthermore, *to* almost always appears between *go* and *football*, or between *go* and events and venues generally.

Multi-word expressions also pose problems for the computational processing of language, as do criteria for decision-making that spans the whole text, not just the current sentence or immediate context. The greatest obstacle to realising such software is the so-called sparse data problem. A vast amount of data is needed to make statements about word choice, including those words which are grammar.

In conclusion

While the future may look rosy, the present is certainly far from grim. A skilful application of currently available software can considerably automate the marking of written work and the internet offers us rich sources of information. At the same time, skilful management of these resources can turn that information into student knowledge.

Acknowledgements

I would like to thank my colleagues at the FI MU, Karel Pala, Pavel Rychly and Lubos Popelinsky, for fruitful discussions and access to much of the software discussed in this article. And to Tom Cobb and Pete Whitelock for their web-based software and their correspondence whilst writing this article, and to Laurence Daudin-Vigot for sharing her experiences with Markin. I would finally like to thank my colleagues and students for being an endless source of data for error analysis!

References

Chapelle, C. A. (1998) *Multimedia CALL: lessons to be learned from research on instructed SLA*. Vol.2, No.1 pp.22-34

Godwin-Jones, B. (2000) "Emerging Technologies, Literacies and Technology Tools/Trends." *Language Learning & Technology*, Vol.4, No.2, pp.11-18.

Krajka, J. (2002) "Correcting student work with the computer - using dedicated software and a word processor." *Teaching English with Technology, A Journal for Teachers of English*, Vol.2, No.4, http://www.iatefl.org.pl/call/j_tech10.htm.

Lewis, M. (1993) *The Lexical Approach*. Hove: LTP.

Sinclair, J.M. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP.