Anything I can do, CPU can do better: a comparison of human and computer grammar correction for L2 writing using BonPatron.com

Terry Nadasdi, University of Alberta
Stefan Sinclair, McMaster University

1. Introduction

The purpose of our paper is to argue that computers can be more efficient than humans when it comes to marking second language texts. This claim is based on the fact that a pedagogically-oriented grammar checking program like *BonPatron.com* can a) identify approximately the same number of errors as human correction; b) do so in a timely manner; c) do so consistently; and d) provide more extensive and/or comprehensible feedback than human marking. Our paper is structured as follows. We will first briefly discuss previous research that has evaluated the efficacy of grammar checkers. We will then present the methodology used in our own study and present a quantitative comparison of human and machine correction. Finally, our results will be discussed and conclusions will be drawn regarding the pedagogical advantages of machine correction.

2. Previous research

The majority of previous studies that have assessed the usefulness of grammar checkers fall into two basic categories: a) those that have evaluated MS Word's grammar checker for Anglophones writing in English; b) those that have considered a variety of grammar checkers for L2 learners of French. We will focus on these latter studies, and simply remark that most evaluations of Word's grammar checker are highly critical since the program generates far too many false positives and does not catch frequent errors[1].

[1] See for example http://papyr.com/hypertextbooks/grammar/gramchek.htm

Many of the studies that have examined French grammar checkers for L2 writing are really software reviews. The research component one finds in such studies is usually quite modest and sometimes problematic[2]. Let us present a brief summary of these studies to give the reader an idea of the general results one finds in previous analyses of grammar checkers for L2 French instruction. Four different software packages have been commonly used in previous studies, these are: *Antidote*, *Correcteur* 101,*The Bilingual Corrector* and *Sans Faute*.

There is no real consensus in terms of whether or not grammar checkers are helpful for second language instruction. Some studies point to a handful of advantages. However, a large number of drawbacks are also reported. Among the advantages, one can mention that, depending on the actual software examined, high rates of error detection are noted (cf. Burston, 2001; Murphy-Judy, 2006) and that some programs provide opportunities for students to reflect on language (Charnet and Panckhurst, 1998; Druel, 2006). Still, many have lamented the fact that grammar checkers suffer from the following deficiencies: a) they generate too many false positives (cf. Jacobs and Rodgers, 1999; Burston, 1998); b) they provide no English interface (cf. Mogilevski, 1998); c) they are not designed for L2 learners (cf. Jacobs and Rodgers, 1999); and they are not pedagogical in orientation (cf. Burston, 1998; Cordier-Gauthier and Dion, 2006; Druel, 2006).

The software used in the present study has not been the object of independent evaluation, but it has been designed in such a way as to address any of the above-mentioned criticisms (see below).

3. Methodology

3.1 The corpus of student texts

[2] For example, the texts are sometimes contrived (e.g.: Jacobs, G. and C. Rodgers, 1999) and the representation of human correction is usually unrealistic (e.g. Cordier-Gauthier and Dion, 2006).

The student compositions used in our study come from submissions to an on-line French grammar checker: bonpatron.com (created by Nadasdi and Sinclair), which currently receives approximately 5000 daily visitors. In early 2006, users were given the option of providing two pieces of demographic information: a) their first language and b) whether they reside in a francophone region, an anglophone region, or a region whether neither English nor French is the majority language. Approximately 5% of users provided this information. Our corpus, then, was made up of the first 30 texts between 1000 and 1500 characters (roughly 250 words), submitted on October 15, 2006, for which a user had indicated that they were an Anglophone residing in a region where English was the majority language. We decided to use 30 texts since it is representative of first year university language courses in Canada. The length of 250 words was chosen since it is typical of such texts.

3.2 Human correction

Previous studies of grammar checkers seem to work under the assumption that there is one infallible human correction against which one should compare grammar checkers. However, the literature on how humans go about marking texts during the course of a school term is minimal. Still, one assumes that there must be a great deal of variability depending on a whole host of variables such as knowledge of the target language, teaching experience, teaching philosophy and other factors such as fatigue, interest, pedagogical focus at that time, etc. While the dearth of research on this topic prevents us from describing how humans go about the correction of L2 texts, we can describe the particular approach used be the authors of this study, both of whom have taught French grammar and composition for many years.

In order to assess the errors in the corpus of texts, we first examined several texts and devised an error legend that could be applied to all the errors we identified. This legend, presented below, is similar to that used by many language teachers:

Error Legend

sp: Spelling (any word caught by spell check); only one error per word coded
agdn: Determiner/Noun agreement error, e.g.: *tes ami*
agan: Adjective/Noun agreement error, e.g.: *premier chose*; past participle agreement, e.g.: *la chose que j'ai remarqué*
agsv: Subject/Verb agreement error, e.g.: *Ils mange*.
mw: Missing word, e.g.: *Je veux _ il mange*.
wf: Wrong word form, such the use of an infinitive instead of past participle, etc.) e.g. *il a manger* instead of *il a mangé*
la: Lexical (anglicism), e.g.: *marcher à l'école*
wc: Word choice (e.g. *avoir* instead of *être*; wrong preposition)
p: Punctuation (capitalization, missing space, missing comma, etc.)
wo: Word order (syntax problems), e.g.: *Je veux la*.
el: Elision, e.g.: *je aime*.

An example of how this legend was applied to the grammatical analysis of an actual text follows:

Example of human corrected text:

Nous avons discuté de ton cas et nous avons des idées que tu devrais regarder si tu veux être en meilleure forme. Le***agdn** premier***agan** chose que nous avons remarqué***agan** a été***wf** le temps***la** que tu te chouches***sp**. Plutôt de***wc** 11h 30, tu devrais couches***wf** à 10h. Ce chagnement***sp**, est le plus importante***agan** de tout. Si tu ***mw** chouches***sp** à 10h, tu pourrais te léver***sp** à une heure matinale. Maintenant**,*p**te***wf** aurais le temps de prendre ton déjeuner. Le***agdn** deuxième chose ***mw** nous avons remaqué***sp*agan** a été***wf** que tu prendsd***p** l'autobus pour arriver à l'école très vite. Nous pensons que tu devrais marcher***la** à l'école. Puis tu auras marcher***wf** pour***la** un kilometer***sp**, tu seras en bonne forme. Probablement, tu n'auras pas le temps ***mw** récontrer***sp** tes amis avant le début des cours ***p**mais ce n'est pas très improtante***sp*agan** …

We first marked the papers individually. Human correction of the texts took about 5 minutes per text. We tried to mark these texts in the same way we had previously marked thousands of texts during our years as language instructors. We took the necessary time, but moved through the texts in the same way anyone would who had to mark 30 of them! Next, we met to discuss each other's corrections, revisit the texts and arrive at a "super human" correction. This is the human correction we used for comparative purposes. In our initial correction, one of the two authors caught 82% of the total errors eventually agreed upon, while the other caught 86%.

3.3. The computer program

The computer program used to evaluate machine correction is an online French grammar checker, www.bonpatron.com, devised by Nadasdi and Sinclair. In order to use the site, a user simply pastes their texts into the main window, clicks the submit button, and receives feedback on common errors. For example, if a user types “le nom de mon mère est Linda et elle et 40 ans”, the site will identify the errors “*mon mère” and “elle *est 40 ans” and provide the following corrective feedback by means of a pop-up window: a) “This noun is feminine. The preceding nouns must also be feminine: e.g.: “Paul loves his mother” = “Paul aime **sa mère**”; b) “when expressing age, the verb *avoir* + *ans* must be used, e.g.: “I am fifteen years old” = *j'****ai*** *quinze ans*.

Errors are placed into three basic categories: a) spelling errors; b) grammatical sequences that are not acceptable; and c) grammatical sequences that are likely unacceptable, for example *je ferrai*, which is acceptable if the intended verb is *ferrer*, but not if it is the more likely *faire*[3].

The 30 texts that make up our corpus were submitted to the website and results were compared with that of the super human correction (it took *BonPatron* less than 1 minute to analyze all 30 texts at once).

It should be pointed out that in our calculations, we considered an error to be identified by *BonPatron* if it met either of two criteria: a) it was flagged immediately or b), in the case of multiple errors within the same structure, it was eventually flagged once the initial feedback was taken into account. To illustrate this, consider the sentence in 1):

1) Le premier chose que nous avons remarqué

When this sentence is submitted to *BonPatron.com,* the program will first identify the most local error (“premier chose”). Once this is corrected, it will catch the other two.

[3] This system of marking resulted in almost no real false positives for *BonPatron*.

Note that there are both computational and pedagogical reasons for proceeding in this fashion.[4]

4. Results

In the following tables, three kinds of results appear: a) Super Human (the combined efforts of both authors); b) Computer1 (these are the results initially arrived at when text were submitted to *BonPatron*; and c) Computer2. This last column reports results for *BonPatron* once we wrote new rules and made adjustments to the rule database, based on results in the Computer1 analysis. It is important to bear in mind that *BonPatron* is a very dynamic project and we continue to add new rules to our database on a regular basis. We have included Computer2 results for two reasons. First, they give a better idea of what the site is capable of and 2) it is a fairer comparison with Super Human correction.

The first series of results we will discuss are presented in Table1:

Table1: General comparative results for human and machine correction

Error type	Super Human	Computer1	Computer2
Grammar	845	564 (68%)	729 (86%)
Punctuation	172	240 (140%)	251 (146%)
Spelling	206	203[5] (99%)	230 (112%)
Elision	15	14 (93%)	17 (113%)
Total	1238	1021 (82%)	1227 (99%)

Results are presented as a comparison with the Super Human correction. For example, the Computer1 correction identified 82% of all errors identified by the humans

[4] Briefly, it seems preferable to not overwhelm the user with complex overlapping error reports, but to have each error clearly identified, explained and corrected in sequence.

[5] The reader might be surprised to see that Computer1 wasn't 100% successful at identifying spelling errors. This is because we classified errors like *fleure* for *fleur* and *vit* for *vite* as spelling errors, rather than word form errors.

(coincidently, that’s the same percentage achieved by one of the authors in their individual correction). We also see that Computer1’s success varies between grammar and the other categories. The lowest success rate is for Computer1 grammar, which still identified almost 70% of the grammatical errors identified by the Super Human correction. This number rises to 86% for Computer2.

Table2 provides a more fined-grained analysis with the grammatical categories broken down according to the codes given in our marking legend:

Table2 : Results according to grammatical subcategories

Error code	Super human	Computer1	Computer2
Adj./Noun agr.	102	62 (61%)	89 (87%)
Det/Noun agr.	116	96 (83%)	113 (97%)
Subj./verb agr.	91	77 (85%)	89 (98%)
Word choice	179	103 (58%)	138 (77%)
Word order	16	7 (44%)	11 (69%)
Missing word	47	25 (53%)	33 (70%)
Lexical anglicism	60	43 (72%)	57 (95%)
Word form	234	151 (65%)	199 (85%)
Total	845	564 (68%)	729 (86%)

We see that both computer corrections have the most difficulty with word order problems and errors related to missing words. Still, these errors are identified in the clear majority of cases and the overall result for all grammatical categories is an impressive 86% for Computer2.

The reader will recall that, as illustrated in Table1, the highest rates obtained for computer correction are for the punctuation category (missing spaces, missing commas, capitalization errors, etc.). In the results we have presented so far, we have considered it an error if the text contained a double punctuation mark (exclamation, question mark, colon or semi-colon) not preceded by a space, as is prescribed by most European style guides. However, Canadian French style guides do not usually espouse this convention. In order to present a more balanced perspective that ignores this stylistic rule, we have revised our results so as to reflect Canadian usage, i.e. the results in Table 3 do not consider absence of a space before a double punctuation mark to be an error:

Table 3: Results when one does not require a space before double punctuation marks:

Error type	**Super Human**	**Computer1**	**Computer2**
Grammar	845	564 (68%)	729 (86%)
Punctuation	91	108 (119%)	119 (131%)
Spelling	206	203 (99%)	230 (112%)
Elision	15	14 (93%)	17 (113%)
Totals	1157	889 (77%)	1095 (95%)

This results in a slight drop in overall error identification rates, but the general results remain the same. In other words, both computer corrections produce much higher rates for punctuation correction than what was produced through human correction.

4.1 Summary of results

This issue of what constitutes an error is a long-debated one and there are various ways of approaching the problem. Still, the results presented in Tables 1-3 make it clear that there is relatively little quantitative difference between computer correction with *BonPatron* and Super Human correction. It is true that humans did identify more grammatical errors. However, as we will see below, there are certain qualitative drawbacks with human correction that greatly diminish any quantitative advantage.

5. Discussion

It is by no means a straightforward task to determine whether or not correction with *BonPatron* is better than human correction. It depends in part on one's goals and one's opinion of what constitutes beneficial pedagogical intervention. However, there are at least two important questions that need to be addressed in any discussion of this type:

a) What does human correction really look like?
b) Do students learn from corrective feedback?

As mentioned, we know relatively little about the way humans mark L2 texts, but it is not reasonable to suggest that this is done in a monolithic fashion. Individual humans mark texts in a variety of manners (in terms of the feedback supplied) and do not consistently identify the same errors. Still, those who have described human L2 correction have tended to view it quite positively; for example:

2) "Enfin, exhaustif dans son repérage … [la médiation humaine] … est pratiquement **infaillible** [our emphasis] même s'il peut parfois être influencé par des facteurs psychologiques (être tolérant ou agacé, se lasser ou même s'ennuyer) ou physiologiques (se fatiguer)" (Cordier-Gauther & Dion, 2003).

This representation of human marking is perhaps true in theory, given the optimum conditions, but this it is far from what happens in reality. These authors are right in drawing attention to the fact that there are a variety of psychological factors that may hinder human correction and which, of course, do not come into play with computers. What we do not know, however, is the **extent** to which they come into play when teachers mark second language texts. Add to this the fact that the linguistic competence of language teachers is highly variable. Many of them are teaching assistants in our universities and graduates with maybe a French minor in our grade school systems. Furthermore, time and energy constraints are also likely to hinder human accuracy. As such, one begins to wonder how close the average teacher really comes to super human correction (aka the team-of-experts-with-time-on-their-hands approach!). This is a research question that needs to be fully investigated, though the task is not an easy one since humans are notorious for behaving differently when they know they are being observed!

It should also be noted that even if the percentages arrived at for human correction in Table1 were realistic, there are still reasons to prefer a computerized pedagogical grammar checker. One question to consider is the following: Is it better to identify a smaller percentage of errors and have the student understand, than it is to identify every single error without pedagogical intervention? This is the real issue. The comparison of human and machine error correction should not be a contest to see who can **identify** the most errors in written language. From a pedagogical perspective, the identification of an error is only a first step to the real goal, which is to teach our students about the grammar of the language they are acquiring. It is our contention then that, in the context of the student – teacher marking experience, humans are at best slightly better at error **identification**, but computers may well be better at error **remediation**.[6] We say this because a program like *BonPatron* is better suited to provide real-time, comprehensible corrective feedback. As we saw above, each error identified by *BonPatron* receives a simple, clear grammatical explanation supported by an example sentence (and the precise morpheme in question is highlighted in bold). Space requirements alone make it impossible for teachers to provide students with rich feedback of this type[7].

The question of feedback in second language research has been the subject of must debate for last ten years. On the one side, we find researchers like Truscott (2004) who claims that people don't learn to write by means of corrective feedback. On the other, we have scholars like Ferris (2004) and Chandler (2003) who maintain that both experimental and theoretical studies suggest that feedback is indeed beneficial for the

[6] Note that we are claiming the superiority of computers in practical terms, though one might suppose that if every student had a dedicated teacher, with infinite patience, and on-call day or night for helping with writing assignments, the balance would be tipped the other way. We also believe that many aspects of grammar and language instruction are best done by humans (in fact, one of the original impetuses of *BonPatron* was to shift the burden of repetitive grammar correction to the computer so that the human instructor could better use the time for other aspects of language learning).

[7] It is true that teachers could use the "insert comment" feature of a program like MS Word, but this does not appear to be the dominant trend in marking since students still tend to submit assignments in paper form.

learning of grammar[8]. In light of this, it should be noted that one important advantage of computerized checkers is that they can provide real-time explanations and supporting examples for every error a student makes. Furthermore, error correction via computer programs like *BonPatron* are interactive and involve student participation since they must interpret the grammatical explanation and make the necessary changes themselves[9]. With human correction, students tend to receive feedback days or even weeks after they have written their text and it is not clear that the proposed revisions are understood or taken into account. The only way human correction can ensure this is if a teacher is able to meet with students and provide clear explanations and examples for each student. However, this is unlikely to happen since there are practical constraints that limit teachers' time and energy.

We also have considerable anecdotal evidence from our users that grammar correction with *BonPatron* does indeed lead to learning. Several telling testimonials are presented in 3):

3) *I am in grade 7 and My French has improved because of Le patron*[10]*, I used to get c-- or c--- even F now i get c+ and sometimes c, and I hope to get a B some day! I am gonna tell EVERY ONE that Le Patron RULES thanks Le Patron you've changed my Life ! Geneviève*

Mon ami m'a montré Le Patron, puis ça m'a sauvé la vie! Mes notes de français (je prends un cours d'histoire en immersion au niveau 10e année) étaient en train de baisser, mais avec l'aide de votre site mon prof a noté beaucoup d'amélioration. Merci! Hannah

Thank you so much Le patron you have helped my mark go up at least 20 percent now I have a happy B+ My teacher was really smart to tell us to use this site! Sarah.

I use your website all the time. It is so awesome. I actually learn a lot of great things and I like seeing that my writing is improving because I get less and less little yellow and red marks. Yeah! I have also recommended LePatron to several of my friends/colleagues! Keep it up, it's really great! Sarah

[8] For an accessible account of this debate and the role of feedback in L2 acquisition, see: http://secondlanguagewriting.com/explorations/Archives/2007/March/ErrorFeedbackinL2Writi_1.html

[9] Carpenter-Binkley, 2002 suggests the same advantage for the *Bilingual Correcteur*.

[10] The *BonPatron* programme was previously referred to as *LePatron*.

Thanks so much for your wonderful service. I am a grade 10 student in Toronto, Ontario, and I just want to take a minute to tell you that your website really did teach me. Alex.

While experimental studies would be necessary to further support such claims, these comments are encouraging and suggest strongly that the feedback provided by *BonPatron* helps students learn and improve their writing skills.

6. Conclusion

The main claim of our study is that a pedagogically-oriented grammar checker like *BonPatron* is at least as good as human correction for helping students improve their L2 writing skills. Our motivation for this brazen claim is that, for all intents and purposes, they produce the same results in terms of error identification and also because a pedagogical grammar checker can provide real-time, comprehensive, corrective feedback, which in term is more conducive to learning. Our results do not suggest that teachers should spend less time teaching students how to write. However, they clearly suggest that they should spend considerably less time marking student compositions. Indeed, they would be better off spending no time on marking, and more time meeting with students on an individual basis to discuss their particular problems. The task of learning a second language is a very difficult one. It is therefore necessary to be judicious in our relegating of teaching tasks. We have argued that grammar correction can be relegated to computers, freeing up time for teachers to ensure their students make greater progress in other areas of linguistic competence.

References

Burston, J. 1999. Software review of Antidote 98. CALICO Journal 16 (2).

Burston, J. 2001. A comparative evaluation of French grammar checkers. CALICO Journal, 13, 2, 3: 104-111.

Carpenter Binkley, S. 2002. Review of The Bilingual Corrector 2.0. Calico Review.

Chandler, J. 2003. The effects of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. Journal of Second Language Writing, 267-296.

Charnet. C. and R. Panckhurst. 1998. Le correcteur grammatical: un auxiliaire efficace pour l'enseignant? Quelques éléments de réflexion.

Cordier-Gauthier, Corinne et Chantal Dion. 2006. La correction et la révision de l'écrit en français langue seconde: médiation humaine, médiation informatique. Alsic. Vol 6, 1: 29-43.

Durel, Patrick. 2006. Utilisation de l'assistant grammatical Antidote dans le cadre d'activités de révision – Analyse exploratoire de protocoles d'observation.

Ferris, D. 2004. The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime...?) . Journal of Second Language Writing, 13, 49-62.

Jacobs, G. and C. Rodgers. 1999. Treacherous Allies: Foreign Language Grammar Checkers. CALICO Journal, 16, 4: 509-530.

Mogilevski, 1998. Le Correcteur 101: a comparative evaluation of 2.2 versus 3.5 Pro in CALICO 16 (20).

Murphy-Judy, Kathryn. 2002. "Sans-Faute Writing Environment," CALICO Software Review.

Truscott, J. 2004. Evidence and conjecture on the effects of correction: A response to Chandler. Journal of Second Language Writing, 13, 337-343.