On the performance of modern digital grammar checkers for native and second language learners

A study of gender marking in French

Marco Bril
Vrije Universiteit Amsterdam

In this study we conducted an experiment aiming to compare the performance of language learners and digital grammar checkers at supplying gender marking in French. A set of test items exhibiting typical gender marking configurations was submitted to three grammar checkers for French (Antidote, Scribens and BonPatron). The outcomes were compared to those of native speakers and second language learners of French at a B1 level.

The results revealed that only Antidote and Scribens outperformed both native speakers and second language learners of French in adjective-noun and fronted noun-past participle agreement constructions. An opposite pattern, however, appeared for clitic-past participle constructions for which native speakers outperformed Antidote and Scribens.

We thus conclude that from the three grammar checkers under investigation, Antidote and Scribens might be effective to improve the native speakers' and second language learners' awareness of gender marking errors, but only in adjective-noun and fronted noun-past participle agreement constructions.

Keywords: L1 acquisition, L2 acquisition, French, gender marking, grammar checkers

1. Introduction

In the last decades, many studies aimed at evaluating the effectiveness and the usefulness of digital grammar checkers in computer-assisted language learning (e.g. Biesemans, 2005; Burston, 1998; Heift & Schulze, 2007; Vernon, 2000). Heift and Schulze (2007) for instance, showed that grammar checkers are very effective in form-focused writing exercises. In agreement with Ellis (2001), form-focused

instruction is a pedagogical practice undertaken by second language teachers by which the students' attention is drawn to language form. In this particular case, form-focused exercises would explicitly target overt grammatical inflection by confronting students with the relevant morphosyntactic rules. Within the context of gender inflection for example, these exercises may consist of fill-in-the-blank sentences in which learners need to fill in the correct gender inflection or sentences containing incorrect gender marking which need to be corrected by the language learners. In this type of writing exercises, grammar checkers can be of help for learners. More precisely, learners can submit their sentences to the grammar checker for correction. The checker, then, highlights potential grammatical errors and provides individual metalinguistic feedback. Based on this feedback, learners can autonomously correct their errors.

The usefulness of grammar checkers in the language learning process of French gender marking may be defined in terms of the number of gender marking errors that the grammar checker is able to check, correct or supply. In this context, checking yields the identification of incorrect grammatical inflections by the checker, without correcting them. Correcting yields both the identification and correction of incorrect grammatical inflections. In contrast to checking and correcting, supplying yields the implementation of grammatical inflection where it is lacking (Burston, 1998, 2008).

This study deals with a comparison between human correctors and grammar checkers when it comes to the supply of gender inflection in French. In this particular category, writing errors are commonly found in L1 and L2 French. Interestingly, the correction rates of French digital grammar checkers are found to differ with regard to particular types of agreement constructions. More specifically, Biesemans (2005) observed that only 21.4% of inflection errors in past participle agreement were corrected by the Antidote grammar checker, in contrast to a correction rate of 62.6% for inflection errors in subject-verb agreement. This suggests that in French, the performance of Antidote is influenced by the syntactic configuration exhibiting gender agreement. More precisely, agreement types hosted in a relatively local syntactic configuration, such as subject-verb agreement, show an enhanced correction performance compared to those in a larger syntactic configuration, such as past participle agreement.

For both native and second language learners of French, the size of the syntactic configuration has also been shown to have an effect on the performance of written inflections (Bril, 2016). In this respect, the size of the syntactic domain can be defined in terms of the number of words intervening between the agreeing elements in the sentence (*Processability Theory*, Pienemann, 1998). As such, adjective-noun constructions (e.g. *la grand**e** voiture* 'the-F.SG. big-F.SG. car-F.SG.') are taken to be hosted in a relatively small syntactic domain, since there are

no intervening elements between the adjective and the noun. In a similar way, fronted noun-past participle constructions (e.g. *la boutique qu'on a trouvée* 'the-F.SG. shop-F.SG. that we have found-F.SG.') can be taken to be hosted in a larger syntactic domain than clitic-past participle constructions (e.g. *je l'ai arrosée, (la plante)* 'I it-F.SG. have watered-F.SG., (the plant-F.SG.)), as there are three words intervening between the noun and the past participle in the first construction and one intervening between the clitic and the past participle in the second construction.

With respect to written language production, Bril (2016) showed that clitic-past participle constructions and noun-past participle constructions trigger more writing errors in gender marking than adjective-noun constructions. Furthermore, noun-past participle constructions were found to trigger more errors than clitic-past participle ones. As in grammar checkers, the difference in writing errors may be related to a difference in the size of the domain: past participle agreement constructions are hosted in a relatively large syntactic domain, whereas adjective constructions exhibit a very local configuration pattern.

2.   Main aim and research question

The main aim of the present paper is to compare the performance of digital grammar checkers and language learners when it comes to gender inflection in written French. This comparison enables us to evaluate the usefulness of these checkers and to formulate implications for the educational practice. We report the results of an experiment investigating the supply of gender marking in different types of gender agreement configurations by three digital grammar checkers (Antidote, Scribens and BonPatron) and by two populations of French language users (native speakers and second language learners). To contrast the performance of grammar checkers and language users, we will use the materials and the language users' results reported in Bril (2016) and compare these results to those of the grammar checkers under investigation in the present study.

We will address the following research question:

Do digital grammar checkers reach a higher performance at supplying gender marking in French than native and second language learners?

To answer this question, we will compare their performance on gender marking in three well-defined gender agreement configurations that differ from each other with respect to the size of their syntactic domain.

The paper is organized as follows: in the third section, three types of grammar checkers and their way of detecting potential syntactic errors will be presented. In this section the main features and performance of the three digital grammar checkers under investigation will also be described. In the fourth section, we will describe the effect of digital grammar checkers on language acquisition. In the fifth section, then, a brief overview of the gender system in French will be presented. In the sixth section we will present the methodology and the results of our study. Finally, in the last section, we will draw conclusions based on these results.

3.   The detection of errors by grammar checkers

In written language production there are roughly two categories of errors: misspellings and syntactic errors. Misspellings can be detected and corrected by a standard spelling checker. In contrast, syntactic errors (e.g. morphosyntactic errors) need to be detected and corrected by a grammar checker. In the present section we will present how grammar checkers detect and correct syntactic writing errors (3.1) and describe the performance of three French grammar checkers under investigation in this study (3.2).

3.1   Types of grammar checkers

To check or correct syntactic errors there is not a universal algorithm by which grammar checkers face these errors. Naber (2003) for instance, distinguishes between three types of algorithms to implement grammar checkers: syntax-based, statistics-based and rule-based. More precisely, syntax-based grammar checkers completely parse the sentences of the text and assign a tree structure to each sentence. Based on morphosyntactic elements (e.g. adjectival or verbal inflection), the checker assigns grammatical features to the individual words of the sentence. The tree structure, then, is used to detect syntactic dependencies between words in this particular sentence. This allows to check the correctness of morphosyntactic elements. In this respect, most grammar checkers are capable of identifying grammatical features of words in one and the same sentence and show difficulties in identifying syntactic errors in long distance dependencies (e.g. Teixeira Martins, Hasegawa, Das Graças Volpe Nunes, Montilha, & Novais de Oliveira, 1998).

Statistics-based grammar checkers rely upon a large corpus containing several types of short sentences (cf. Chodorow & Leacock, 2000). Based on frequent sequences of words in the corpus, similar sequences in the text to be checked, can be considered as correct. However, correct sequences in this particular text being uncommon in the corpus may be considered as errors in the text.

In a similar way, rule-based grammar checkers detect potential errors by comparing sentences to frequent sequences in a corpus. Based on the corpus, software developers manually program grammatical rules for the particular checker. Rule-based checkers, then, apply these relevant grammatical rules to the text to be checked (cf. Park, Palmer, & Washburn, 1997).

Most grammar checkers have been shown to rely on a combination of the syntax-based and statistics/rule-based approach to detect and correct writing errors (Jensen, 1993). Since the types of errors made by native speakers vs. second language learners differ with respect to the frequency in language production, grammar checkers use learning corpora which are specific for these particular populations of language users.

3.2   French grammar checkers: Antidote, BonPatron and Scribens

Antidote, BonPatron and Scribens are well-known digital grammar checkers for French. More specifically, Antidote is a grammar and spell checker which has been developed by the Canadian software enterprise Druide Informatique for both native speakers and second language learners of French. It can be used as a supplementary grammar and spell corrector and can be applied to all levels of French language acquisition. For texts submitted to Antidote, the user can select the types of errors which the program needs to correct. As such, the selected types of errors are marked by using underlines in colors marking particular categories of writing errors (e.g. the red colored underline marks writing errors in the morphosyntactic domain). The user can accept or ignore the correction proposed by the Antidote program. Antidote also provides corrective feedback to the user explaining why the underlined item has been written incorrectly. More specifically, the appropriate grammatical rule along with an example pops up when the user clicks on the underlined item.

In a similar way, BonPatron is mainly used to correct errors in texts written by second language learners of French. This digital grammar checker has been developed at the Canadian University of Alberta and McMaster University for English learners of French and is available online (bonpatron.com). Here also, particular categories of writing errors are distinctively color-marked. In BonPatron, errors are surrounded by a colored text box over which the user can pass the cursor activating feedback concerning the grammatical rule that has been violated by the language learner. In contrast to Antidote, BonPatron has a certain degree of self-learning capacity: It collects all text data submitted to the checker to improve its own correction algorithms. An additional feature that may be of interest to second language users is the summary of signaled errors provided at

the end of the text correction. This may contribute to the users' awareness of particular writing errors.

For the digital grammar and spell checker Scribens no peculiarities need to be mentioned. Scribens corrects writing errors and provides corrective feedback on grammatical errors in roughly the same way as Antidote and BonPatron. The Scribens program is available online (scribens.fr) and has been developed for native speakers of French by a private software developer in France.

Concerning the performance at correcting inflections errors, we already mentioned that Antidote is not able to correct all morphosyntactic errors. In a similar vein, BonPatron corrects 88% of the morphosyntactic errors found in the written productions of L2 French language learners (Burston, 2008).

Nadasdi and Sinclair (2007) provided a more detailed analysis of morphosyntactic errors that are corrected by BonPatron. They submitted 30 texts to the grammar checker that were written by English learners of French. These learners were enrolled in the first year of a French language course in a Canadian university and were asked to write a text of roughly 250 words. The writing errors that were presented to BonPatron were divided into four stylistic categories: grammar, punctuation, spelling and elision. In parallel, the same writing errors were presented to human correctors. The correction performance of BonPatron was computed in terms of the percentage of errors identified by human correctors. The results revealed that in grammar, spelling and elision, human correctors identified more errors than BonPatron. Contrastingly, BonPatron identified more errors in punctuation than human correctors. All errors identified by human correctors and BonPatron were actual errors and no incorrect identifications. Based on these findings, the developers made adjustments to the rule database of BonPatron. A dataset of written texts was then submitted to the new version of the grammar checker. With respect to grammar correction, an in-depth analysis shows that the correction performance was lowest in adjective-noun agreement constructions (i.e. only 61% of the errors identified by human correctors was corrected in version 1 and 87% in version 2), followed by subject-verb agreement constructions (85% in version 1 and 98% in version 2), and by determiner-noun agreement constructions (87% in version 1 and 97% in version 2). Based on these data, the authors concluded that the performance level of the enhanced version of BonPatron reaches the performance level of human correctors in the correction of inflections.

To the best of our knowledge, for the other two French grammar checkers (Scribens and Antidote), an in-depth analysis on digital correction performance seems to be lacking altogether.

4.   The effect of digital writing programs on language acquisition

In line with Heift and Schulze (2007), Zhao (2003) found overall positive effects of digital technology on language acquisition. More specifically, the author analyzed nine studies in a meta-analysis aiming to explore the overall effectiveness of digital technology in language acquisition. These particular studies compared learners' performance in writing, speaking, listening, reading and cultural knowledge in settings which were controlled for two different types of input: input of a teacher vs. input by means of digital technology. The results of the meta-analysis revealed that digital technology has an overall positive effect on language acquisition. However, the author remarked that the number of purely experimental studies on digital technology in language learning is quite limited. Moreover, the sample sizes of the studies included in the meta-analysis, are small.

In a similar vein, the use of digital writing programs has been shown to be more effective than traditional instruction when it comes to the written production of native speakers. Bangert-Drowns (1993) for instance, carried out a meta-analysis on 32 studies focusing on the use of digital writing aids, such as grammar or spell checkers, vs. traditional writing instruction. More precisely, all studies compared two groups of native speakers which both received the same writing instruction. However, one group was allowed to use a spell or grammar checker for writing assignments, while the other group received corrections from the teacher. The overall results showed that native speakers using digital writing aids attained a higher quality of their writing productions as compared to those who received corrections from the teacher or peers. Furthermore, the written documents of native speakers using writing aids were found to be longer than those written by native speakers who received no digital feedback.

With respect to second language learners, a positive effect of digital programs on writing performance has also been shown. Chuo (2007) for example, tested the potential effect of the WebQuest Writing Instruction Program (WQWI) on writing performance in Taiwanese learners of English. These learners were second year students of the English language minor or major in a junior college division in Taiwan. By means of a pretest-posttest design the participants were randomly selected for the intervention group ($n=54$) or the control group ($n=54$). The intervention group completed a set of writing tasks with the WQWI program, while the control group completed the same set of tasks printed on handouts and was in the traditional classroom setting. Both groups were firstly assessed by means of a writing performance pretest. After a fourteen weeks intervention period a writing performance posttest was administered in both groups. Comparisons between both groups revealed that the intervention group reached a higher score in the posttest than the control group. Furthermore, the scores

on the pretest significantly correlated to those obtained on the posttest. Based on these findings, the author concluded that the use of WQWI is more effective than a traditional writing setting in second language writing instruction.

A similar effect of computerized programs has been shown by Wang, Shang and Briody (2013). More specifically, the authors focused on the potential effect of Automated Writing Evaluation (AWE) on the written production of Taiwanese learners of English ($n = 57$). In the same experimental design as Chuo (2007) the participants were assessed in terms of writing accuracy, learner autonomy and interaction with other language learners. The group of participants in which AWE was used achieved a higher score in the writing test and a higher score on learner autonomy than the group of participants which did not use AWE. These results clearly show that computerized programs are very effective to enhance the accuracy of written language production.

However, as shown by Biesemans (2005), digital grammar checkers show difficulties when it comes to the correction of some particular grammatical errors. As such, these difficulties can negatively affect the written productions of language users. More precisely, Jacobs and Rodgers (1999) asked two groups of second-year university students who were learning French as a second language, to correct four texts containing grammatical, lexical and orthographical errors. These texts were corrected by the participants in the first group who had access to digital grammar and spell checkers, while the same texts were corrected with handbooks in the second group. Interestingly, the first group corrected gender and orthographical errors more accurately than the second group, but grammatical and lexical errors less accurately than the second group. This observation may be explained by the fact that digital checkers show difficulties in the correction of particular grammatical and lexical errors (e.g. Burston, 2008). The performance of language users at correcting writing errors, thus, might be influenced by the correction rate of digital checkers. The reason for this correlation can be related to the fact that a positive experience with computers triggers strong confidence in their decisions (Gueutal, 1989).

5.   Gender marking in French

The present study will use French gender marking as test bed to compare the performance level of BonPatron, Scribens and Antidote to that of L1 and L2 learners of French. For that purpose, we will first provide a brief overview of the gender system in the French language. Firstly, the language exhibits a two-gender system with masculine and feminine gender marking. Gender marking is present

on definite and indefinite articles (see (1a) and (1b) for masculine and feminine respectively), on adjectives and on past participles.

(1) a. Le / Un   grand**ø**   cadeau
          The-M.SG. / A-M.SG. big-M.SG. present-M.SG.
          'The / A big present'
     b. La / Une   grand**e**   voiture
          The-F.SG. / A-F.SG. big-F.SG. car-F.SG.
          'The / A big car'

Masculine gender is not overtly expressed on regularly inflected non-derived adjectives[1] (see (1a)). The feminine gender, however, exhibits overt gender marking (i.e. + *e* on the adjective), as in (1b).

Secondly, the past participle agrees with the direct object taking the form of an object clitic (see (2)) or a noun (see (3)). To trigger overt gender (and number) agreement, the direct object must precede the past participle. Here also, masculine gender is not overtly expressed ((2a) and (3a)), whereas feminine gender marking is overtly expressed by an *e* ending on the past participle ((2b) and (3b)).

(2) a. Je l'   ai   arrosé**ø**,   (l'arbre)
          I it-M.SG. have watered-M.SG. (the tree-M.SG.)
          'I watered the tree'
     b. Je l'   ai   arrosé**e**,   (la plante)
          I it-F.SG. have watered-F.SG. (the plant-F.SG.)
          'I watered the plant'

(3) a. Le   ballon   qu' on a   trouvé**ø**
          The-M.SG. ball-M.SG. that we have found-M.SG.
          'The ball (that) we found'
     b. La   boutique qu' on a   trouvé**e**
          The-F.SG. shop-F.SG. that we have found-F.SG.
          'The shop (that) we found'

Concerning the acquisition of gender inflection in written language production, Bril (2016) showed that in both native speakers and second language learners of French, the written accuracy of this type of inflection is influenced by the size of the syntactic domain in which the agreement configuration is hosted. More precisely, the larger the syntactic domain, the more difficult to acquire the gender inflection. Against this background, the author demonstrated that gender marking in adjective – noun constructions is applied more accurately than in clitic –

---

1. In derived adjectives (e.g. *menteur*-M.SG. vs. *menteuse*-F.SG. 'lying') the masculine gender is overtly expressed by the suffix bearing the +M gender feature.

past participle and fronted noun – past participle constructions. The reason for this can be found in the fact that adjective – noun agreement takes place in a smaller syntactic domain than the later ones. Furthermore, gender marking in clitic – past participle constructions is applied more accurately than in fronted noun – past participles ones.

6. The experiment

To answer the research question, we compared the rates of supplying gender marking in three digital grammar checkers for French to those in native and second language learners of French reported in Bril (2016). The task for the grammar checkers consisted of the same fill-in-the-gap task which was administered with human learners in Bril (2016). The rates of supplying gender marking in native speakers and second language learners of French were taken from data that were collected for the Bril (2016) study as well.[2]

6.1 Method

6.1.1 *Participants in Bril (2016)*

The participants tested in Bril (2016) were monolingual native speakers of French ($n = 28$; age range = 14–15 years) from the west of France. Each participant had to confirm that the home language was French and that he/she had no language disorder, such as dyslexia. The Dutch learners of French ($n = 26$; age range = 17–18 years; proficiency = B1 level of the common European framework of reference) attended courses in French language for 5 years at the highest level of the Dutch secondary school system (i.e. *voorbereidend wetenschappelijk onderwijs* 'academic university preparatory education'). Participants did not have any known language

---

2. We are aware of the fact that Bril (2016) focussed on language *production*, while digital grammar checkers *monitor* linguistic data (i.e. check, correct or supply). Within the context of human language users, it is well-known that language production and language monitoring are different processes. Therefore, production data and monitoring data of humans are not comparable. The digital grammar checkers tested in this study, however, only supply linguistic data. In such a process grammatical rules are implemented by means of algorithms which are dependent upon the grammatical context. Against this background, the accuracy of correction is highly related to the implementation of grammatical rules (cf. Burston, 2008).

With respect to humans, the accuracy of language production is also highly related to the correct implementation of grammatical rules. Under such a view, we consider the process of supplying gender marking by digital checkers and humans as comparable in this study.

disorder, such as dyslexia. Both the native speakers and second language learners received explicit grammar instructions in their French language courses.

6.1.2 *Materials*

To allow for comparison, the text stimuli that were presented to the digital grammar checkers were identical to those used with human language learners in Bril (2016). More precisely, the test stimuli taken from Bril (2016) were presented as a fill-in-the-gap task requiring the learner or the digital grammar checker to supply overt feminine gender marking on adjectives or past participles when deemed necessary. The test items of both the native speakers' and second language learners' task were controlled for three types of gender agreement constructions representing different sizes of syntactic domains: with attributive adjectives, with past participles preceded by an object clitic and with past participles preceded by an object noun.

Native speakers' task
Each test condition contained 40 test items which were all in a feminine singular context. To avoid mistakes with respect to the lexical gender of the noun or the clitic, the gender was overtly expressed in all conditions. Furthermore, 30 filler items were added in order to verify whether the participants were able to comply with the test requirements. Only the results of the test items were included in the analysis.

Second language learners' task
A subset of gender marking contexts was further used for in-depth analysis. In this stimuli set only contexts with feminine singular marking were further taken up for analysis ($n=46$). Test items targeting plural and/or masculine agreement ($n=104$), were considered as filler items and were not taken into consideration any further.

From a total of 46 items, 5 test items which represented the adjective-noun agreement condition (4), 7 test items exhibited the clitic-past participle agreement condition (5) and 34 test items represented the fronted noun-past participle agreement condition (6).

(4)   Attributive adjective
      La   meilleur_____ joueuse   a   été   sélectionnée pour la
      The-F.SG. best-F.SG.   player-F.SG. has been selected   for the
      compétition
      competition.
      'The best player has been selected for the competition'

(5)   Past participle with object clitic
      La   plante   n'avait plus d'eau.
      The-F.SG. plant-F.SG. did not have water.
      Je l'   ai   arrosé_____
      I   it-F.SG. have watered-F.SG.
      'The plant did not have water. I watered it.'

(6)   Past participle with fronted noun
      C' est la   fleur   exotique   que j' ai   vu____   dans le   jardin
      It is   the-F.SG. flower-F.SG. exotic-F.SG. that I have seen-F.SG. in   the garden
      'It's the exotic flower (that) I saw in the garden'

6.1.3   *Procedure*

The participants tested in Bril (2016) were asked to write down on paper the correct gender morpheme in each test item if needed. The experimental setting of the native speakers and second language learners was in a class room at a secondary school in the west of France and in the west of The Netherlands respectively.

The task was done individually and under supervision of the teacher. Furthermore, the presentation order of the test items was counter-balanced in three versions. All participants finished the task well within the 50 minutes.

The outcomes from the digital grammar checkers were obtained by feeding the native speakers' test items taken from Bril (2016) to Antidote and Scribens for correction. Similarly, the test items of the second language learners' test were submitted to (the updated version of) BonPatron (cf. Nadasdi & Sinclair, 2007) and Antidote. The reason why Scribens was specifically tested with the native speakers' test items, BonPatron with the second language learners' test items and Antidote with both tasks, can be found in the fact that Scribens was particularly developed for native speakers of French, BonPatron for second language learners of French and Antidote for both populations (see Section 3.2). Here, the filler items were removed as there was no need to control for the task in a digital experiment environment.

6.1.4   *Analysis*

For both the native and second language learners, correction rates were computed per test condition (Bril, 2016). These rates were expressed in terms of the percentage of correct gender inflections which have been supplied. For the digital grammar checkers, we computed the rates of supplying gender marking per test condition. These rates were expressed in terms of the percentage of correctly supplied gender inflection. Subsequently, we compared the rates of language users and digital checkers in order to test whether there are significant differences with

respect to the performances. In this respect, the rates of supply of the grammar checkers were taken as the test values to which those of the language users were contrasted. For parametric data we conducted a one-sample *t*-test and for non-parametric data a one-sample Wilcoxon test.

6.2   Results

In Table 1 an overview of the rates of supplying gender marking by grammar checkers is presented.

Table 1. Percentages of correctly supplied gender marking per condition

Type of agreement	Antidote L1 test items	Scribens L1 test items	BonPatron L2 test items	Antidote L2 test items
Adjective	100	100	75	100
Object clitic	40	0	57	43
Fronted noun	98	95	53	100

For both groups of French language users, we will present the rates of correctly supplying gender marking taken from Bril (2016). The data set of rates found in the native speakers' group revealed to be non-parametric and was therefore reported by means of 5 parameter statistics (Table 2). In the second language learners' group the data set revealed to be parametric. Therefore, the means and standard deviations were reported (Table 3).

Table 2. Percentages of correctly supplied gender inflection in native speakers of French

Type of agreement	Minimum	Lower quartile	Median	Upper quartile	Maximum
Adjective	75	85	95	100	100
Object clitic	0	42.50	95	100	100
Fronted noun	3	45	76.50	85.75	100

Table 3. Percentages of correctly supplied gender inflection in second language learners of French

Type of agreement	Mean	SD
Adjective	81	22
Object clitic	54	30
Fronted noun	45	34

To allow for comparison between the grammar checkers and human language users, the results are visualized in Figure 1 (native speakers, Antidote, Scribens) and Figure 2 (second language learners, Antidote, BonPatron).
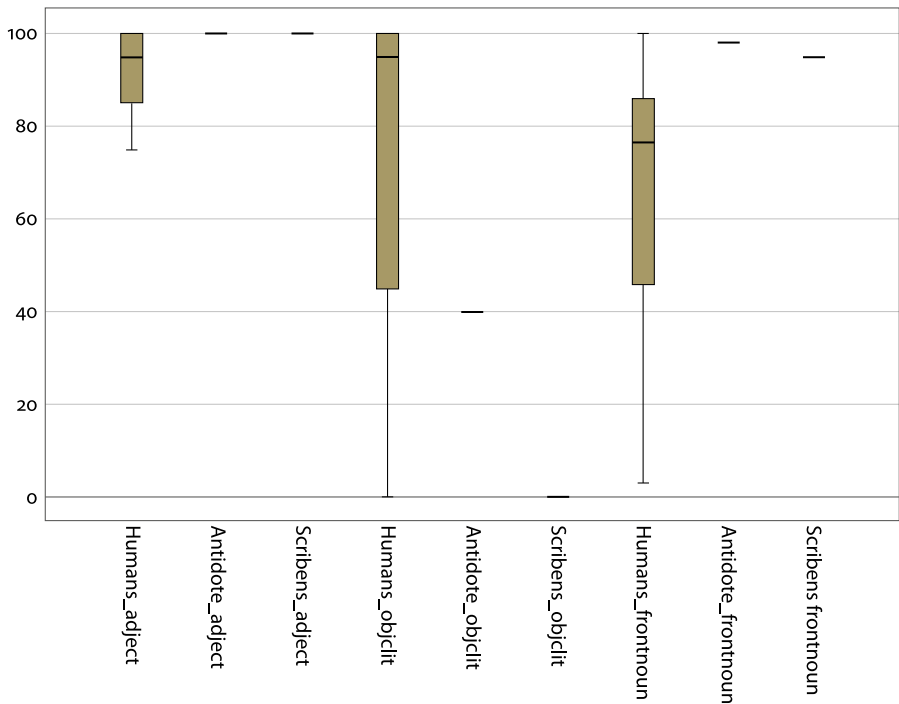

Figure 1. Overview of correctly supplied gender marking in native speakers of French, Antidote and Scribens in %

To compare the rates of supply of the grammar checkers to those of the French language users, we conducted a one-sample Wilcoxon test for the L1 grammar checkers and a one-sample t-test for the L2 grammar checkers on each of the agreement constructions. For all statistical analyses the α level of significance was set at .05.

With respect to the L1 grammar checkers (i.e. Antidote and Scribens), significant contrasts for all types of agreement constructions are observed between these grammar checkers and native speakers of French. More specifically, Antidote performs better in supplying gender marking than native speakers both in adjective-noun agreement constructions ($z(28) = -3.54$; $p < .001$) and in fronted noun-past participle agreement constructions ($z(28) = -4.47$; $p < .001$). In clitic-past participle agreement constructions, however, native speakers outperform Antidote ($z(28) = 3.83$; $p < .001$).
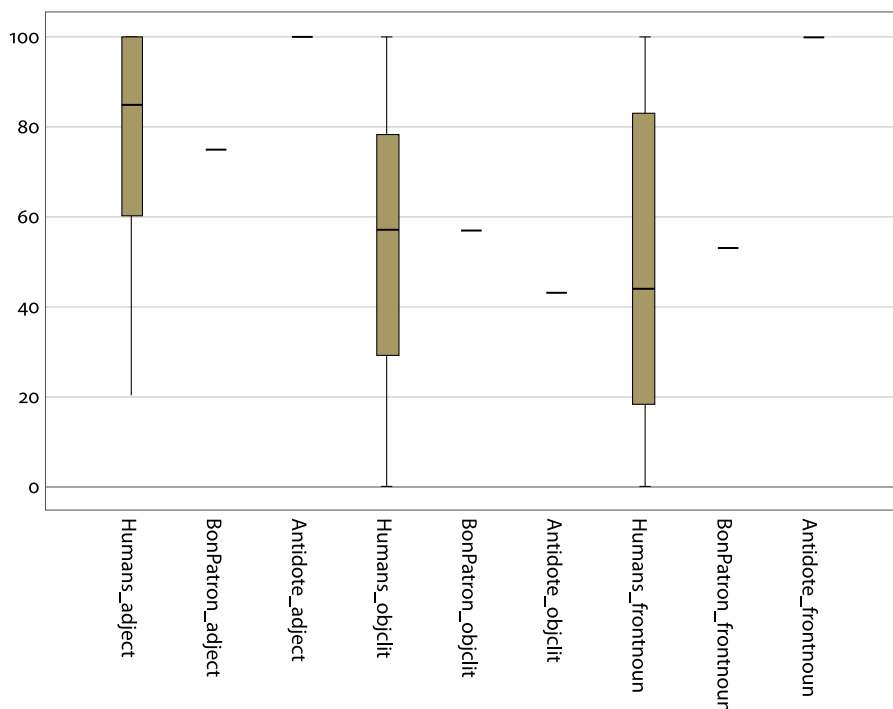
Figure 2. Overview of correctly supplied gender marking in second language learners of French, Antidote and BonPatron in %

Similar contrasts in correctly supplying gender marking are found between Scribens and native French language users. More precisely, Scribens performs better than native speakers both in adjective-noun agreement constructions ($z(28)=-3.54$; $p<.001$) and in fronted noun-past participle agreement constructions ($z(28)=-4.44$; $p<.001$). In clitic-past participle agreement constructions, however, native speakers outperform Scribens ($z(28)=4.57$; $p<.001$).

With respect to the L2 grammar checkers (i.e. BonPatron and Antidote), no significant contrasts are observed between BonPatron and second language learners of French (i.e. adjective-noun agreement constructions: $t(25)=1.40$; $p=.172$; clitic-past participle agreement constructions: $t(25)=-.44$; $p=.664$; fronted noun-past participle agreement constructions: $t(25)=-1.19$; $p=.244$). Concerning Antidote, significant contrasts between this grammar checker and second language learners of French are revealed for adjective-noun agreement constructions and fronted noun-past participle agreement constructions. More specifically, Antidote reaches a higher performance at supplying gender marking than second language learners in both adjective-noun agreement constructions ($t(25)=-4.46$; $p<.001$) and fronted noun-past participle agreement constructions ($t(25)=-8.18$; $p<.001$).

The supply of gender marking in clitic-past participle agreement constructions, however, was not significantly different between Antidote and second language learners ($t(25) = 1.93$; $p = .066$).

6.3   Discussion

The results of this study show that grammar checkers reach a higher performance at supplying gender marking than native speakers and second language learners of French, except for clitic-past participle agreement constructions. The reason for this finding can be sought in the fact that gender marking has been shown to remain problematic in the written production of advanced native and second language learners of French (e.g. Bril, 2018; Bartning, 2000). As such, the advanced language users (both the natives and the learners) tested in the Bril (2016) experiment may be unable to outperform grammar checkers when it comes to this particular grammatical phenomenon. Besides the effect of the size of agreement configurations on the accuracy of written gender marking, the results, thus, may be explained by the difficulty of the acquisition of gender marking amongst advanced native and second language learners of French.

Contrastingly to the performance of human language users, Antidote and Scribens performed at ceiling in supplying gender marking in adjective-noun constructions and fronted noun-past participle constructions. With respect to the latter type of agreement construction, Biesemans (2005) found a significantly lower correction rate (i.e. 21%) for Antidote. The reason for this may be sought in the fact that the outcomes of this study did not differentiate between different types of past participle agreement constructions. In the present study, however, the rates of supply were specified per type of past participle agreement construction. This revealed that the rate of supply differs with respect to the type of this particular agreement construction. More specifically, gender inflection was found to be supplied correctly more often in fronted noun-past participle constructions than in clitic-past participle constructions.

An interesting observation is the fact that all grammar checkers tested in this study seem to struggle with clitic-past participle agreement constructions. For this condition only, native French language users performed better than digital grammar checkers. The difficulty for grammar checkers to detect errors in this particular type of agreement constructions may be related to the algorithm underlying the parsing of sentences. As shown by Teixeira Martins et al. (1998), most grammar checkers are capable to identify potential errors in one and the same clause. Since the agreement relation in clitic-past participle constructions is established between the clitic pronoun carrying the same grammatical features as its referent in the previous clause, and the past participle in French, grammar checkers

need to identify the grammatical feature of the referent which is hosted outside the checking domain of the checkers. This may complicate the detection (and correction) of writing errors in clitic-past participle constructions. Against this background, the effect of the size of the syntactic domain seems to negatively affect the performance of digital grammar checkers in clitic – past participle constructions.

Our results may have important implications for the educational practice, especially with respect to second language teaching. The accuracy of gender marking by French grammar checkers is clearly not uniform across syntactic constructions and grammar checkers. Teachers using a grammar checker in form-focused writing instructions may need to decide on which grammar checker to use depending on the particular grammatical condition they are teaching and on the target population. For native speakers of French, Antidote and Scribens are very effective when it comes to supplying written gender marking in adjective-noun and fronted noun-past participle agreement constructions. Within this context, native speakers receive corrective feedback to reflect and edit their written productions (cf. Vernon, 2000). Based on Bangert-Drowns (1993), Chuo (2007) and Wang, Shang and Briody (2013), both Antidote and Scribens might be taken as useful to enhance the written accuracy of these particular grammatical constructions. However, these very same grammar checkers might not be as effective when it comes to teaching gender marking in clitic-past participle agreement constructions. In agreement with Jacobs and Rodgers (1999), the shortcomings in the correction of clitic-past participle agreement by these checkers may even negatively influence native speakers' performance in this grammatical context.

Concerning grammar checkers developed for second language learners of French, all grammar checkers revealed to be less effective than those developed for native speakers. Specifically, BonPatron did not do any better than second language learners themselves on gender agreement. Therefore, BonPatron does not provide the appropriate corrective feedback to improve the learners' awareness of gender marking. In a similar vein, Antidote does not provide to second language learners the appropriate corrections when it comes to gender agreement in clitic-past participle agreement constructions. Yet, to improve second language learners' awareness of written gender marking in adjective-noun and fronted noun-past participle agreement constructions, Antidote might be effective (cf. Jacobs & Rodgers, 1999). It is noteworthy that these results are found in second language learners at a B1 level of proficiency and that this particular population received explicit grammar instruction on inflectional morphology. Further research needs to be done to investigate whether the effectiveness of these particular checkers in second language writing instruction differs with respect to the proficiency level of the language learners.

In addition to the fact that both language teachers and students need to decide on which grammar checker to use to improve written accuracy, Chen & Cheng (2008) recommended to use digital checkers to correct early drafts of language learners, followed by human feedback for the improved version of their written productions. The authors evaluated the experience of learners who worked with Automated Writing Evaluation (AWE) programs and found that they were frustrated when their drafts were only corrected by AWE programs. Consequently, this frustration limited their writing performance. However, this was not the case when the early drafts were corrected by AWE programs and the improved version of these particular drafts were corrected by human correctors. These evaluations, thus, revealed that the implementation of digital checkers in the educational practice might be most effective when the early drafts are corrected by these checkers, followed by human correctors. Further research needs to be done to investigate whether this also holds for the accuracy of inflectional morphology in the written production of the language users under investigation in this study.

7.   Conclusion

In this study we conducted an experiment aiming to compare the performance of human language users (advanced native and second language learners) and digital grammar checkers at supplying gender marking in French. Here, we focused on adjective-noun, clitic-past participle and fronted noun-past participle agreement constructions. Our results showed that grammar checkers developed for native speakers of French (i.e. Antidote and Scribens), outperformed native speakers of French in adjective-noun and fronted noun-past participle agreement constructions. However, in clitic-past participle agreement constructions, digital grammar checkers could not be of assistance to native speakers of French.

With respect to grammar checkers developed for second language learners of French (i.e. Antidote and BonPatron), no difference between BonPatron and second language learners at a B1 level was observed with respect to the correct supply of gender marking in all three types of gender agreement constructions tested in our experiment. Antidote, however, did reach a higher performance than this population of second language learners in adjective-noun and fronted noun-past participle agreement constructions, but not in clitic-past participle agreement constructions.

We may thus conclude that Antidote might be of help to native speakers and second language learners at a B1 level and Scribens to native speakers by improving their awareness of written gender marking in adjective-noun and fronted

noun-past participle agreement constructions. BonPatron is clearly less effective for this purpose.

References

Antidote (versienummer 10): Digitale spelling- en grammaticachecker voor Engels en Frans. Montréal: Druide Informatique. https://www.antidote.info/en/antidote-10/french-module

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational research*, *63*(1), 69–93. https://doi.org/10.3102/00346543063001069

Bartning, I. (2000). Gender agreement in L2 French: Pre-advanced vs. advanced learners. *Studia Linguistica*, *54*(2), 225–237. https://doi.org/10.1111/1467-9582.00062

Biesemans, K. (2005). Les correcteurs d'orthographe et leur utilité didactique: Analyse de productions d'apprenants de FLE. Unpublished master's thesis, K.U. Leuven.

BonPatron: Online spelling- en grammaticachecker voor Frans. Alberta: Nadaclair Language Technologies inc. https://bonpatron.com/

Bril, M. (2016). Syntactic complexity and inflections in the written production of L1 and L2 French. *Bucharest Working Papers in Linguistics*, *XVIII* (2), 99–114.

Bril, M. (2018). *Persistent grammatical writing errors of L1 and L2 learners of French: Analysis and remedy*. Amsterdam: VU University Press.

Burston, J. (1998). Review of Antidote 98. *CALICO Journal*, *16*(2), 197–212. https://doi.org/10.1558/cj.34912

Burston, J. (2008). BonPatron: An online, spelling, grammar and expression checker. *CALICO Journal*, *25*(2), 337–347.

Chen, C. F. E., & Cheng, W. Y. E. C. (2008). Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94–112.

Chodorow, M., & Leacock, C. (2000). *An unsupervised method for detecting grammatical errors*. Paper presented at the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, USA.

Chuo, T. W. I. (2007). The effects of the WebQuest Writing Instruction Program on EFL learners' writing performance, writing apprehension and perception. *TESL-EJ*, *11*(3). Available from http://tesl-ej.org/ej43/a3.pdf

Ellis, R. (2001). Investigating form-focused instruction In R. Ellis (Ed.), *Form-focused instruction and second language learning* (pp. 1–46). Malden, MA: Blackwell.

Gueutal, H. G. (1989). Utilizing high technology: Computer-aided design and user performance. *Information and Management 17*(1), 13–21. https://doi.org/10.1016/0378-7206(89)90051-7

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning*. New York, Routeledge.

Jacobs, G., & Rodgers, C. (1999). Treacherous allies: Foreign language grammar checkers. *Calico Journal*, *16*(4), 509–530. https://doi.org/10.1558/cj.v16i4.509-529

Jensen, K. (1993). PEG: the PLNLP English Grammar. In K. Jensen, G. E. Heidorn, & S. D. Richardson (Eds.), *Natural Language Processing: The PLNLP Approach* (pp. 29–45). Kluwer Academic Publishers, Boston. https://doi.org/10.1007/978-1-4615-3170-8_3

Naber, D. (2003). A rule-based style and grammar checker. Unpublished doctoral dissertation, University of Bielefeld, Germany.

Nadasdi, T. & Sinclair, S. (2007). Anything I can do, CPU can do better: A comparison of human and computer grammar correction for L2 writing using Bonpatron.com. Unpublished manuscript. Retrieved Sept, 1, 2010.

Park, J. C., Palmer, M., & Washburn, G. (1997). An English grammar checker as a writing aid for students of English as a second language. *Proceedings of the Conference of Applied Natural Language Processing (ANLP)*, Washington, DC. https://doi.org/10.3115/974281.974296

Pienemann, M. (1998). *Language processing and second language development: Processablity theory*. Amsterdam, John Benjamins Publishing. https://doi.org/10.1075/sibil.15

Scribens: Online spelling- en grammaticachecker voor Frans. Montpellier: Thiebaut Alban. https://www.scribens.fr/

Teixeira Martins, R., Hasegawa, R., Graças Volpe Nunes, M. das, Montilha, G., & Novais de Oliveira, O. (1998). Linguistic issues in the development of ReGra: A grammar checker for Brazilian Portuguese. *Natural Language Engineering*, *4*(4), 287–307. https://doi.org/10.1017/S135132499800206X

Vernon, A. (2000). Computerized grammar checkers 2000: Capabilities, limitations, and pedagogical possibilities. *Computers and Composition*, *20*(3), 329–349. https://doi.org/10.1016/S8755-4615(00)00038-4

Wang, Y. J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *Calico Journal*, *21*(1), 7–27. https://doi.org/10.1558/cj.v21i1.7-27

Address for correspondence

Marco Bril
Vrije Universiteit Amsterdam
Faculteit der Geesteswetenschappen
Taal, Literatuur en Communicatie
De Boelelaan 1105
1081 HV Amsterdam

m.bril@vu.nl
https://orcid.org/0000-0001-9114-0321

Publication history