The Treatment of Errors Made by French Second Language Learners in The Use of Object Clitic Pronouns through The Use of a Fine-Tuned Deep Learning Model

Adel Jebali
Département d’études françaises
Concordia University
Montreal, Canada
Email: adel.jebali@concordia.ca

***Abstract*— Object clitic pronouns (particularly third person pronouns) in French are problematic for second and foreign language learners. As a result, several researchers, such as [1], have observed that French second language (L2) learners frequently use avoidance strategies to avoid using these forms, even when doing so allows them to lighten their discourse (written or oral) by avoiding repetition. This is one of the reasons we were interested in technological tools that could assist these learners in comprehending these clitics. We therefore conducted a study with a tripartite goal: to uncover a corpus of L2 French productions focusing on clitics, to use this corpus to train a state-of-the-art deep learning model (CamemBERT), and to implement the trained model to detect learners' errors when producing the forms under study. This model was found to be over 99% reliable when tested. Furthermore, when evaluated on sentences with different turns of phrase than those encountered during training, the model detects errors with the same degree of reliability. This model constitutes a significant advancement in the automatic processing of interlanguage and can be used to develop tools for French L2 learners.**

Keywords*— *French L2; Object pronoun clitics; deep learning; CamemBERT; model.

I. INTRODUCTION

The French Object Clitics (OCs) are primarily personal or oblique pronouns found in contexts, such as in sentence (1b):

(1) a. *Marie a lu la lettre*.
b. *Marie l’a lue*.

In sentence (1b), the OC pronoun *la* (elided as *l’*) is a prosodically weak element that precedes and attaches to the verb while having a syntactic function as the verb’s object. This differs from the behavior of equivalent pronouns in English, for example. In the example (2b), the pronoun *it* is placed after the verb, in the same position occupied by the Nominal Phrase (NP) which it replaces, *the letter*.

(2) a. Mary read the letter.
b. Mary read it.

This peculiar behavior of OCs in French, along with other particularities of these elements, causes confusion among the learners of this language. These learners, therefore, resort to several strategies to compensate for their lack of mastery of these forms. Some authors, such as [1], have raised the issue of omission and avoidance, while [2] also highlights other strategies, such as the repetition of the NP. In addition, these learners generally make errors in positioning the OC relative to the verb and the auxiliary, or make false agreements in gender, number or person with the antecedent; in addition to the grammatical case errors discussed in [3]: using the accusative instead of the dative and vice versa.

All these strategies and errors are well represented in the authentic corpus that we used to fine-tune a deep learning model aimed at classifying French L2 learner’s productions into three categories, as explained in the following sections.

In Section II of this paper, The data and corpus used in this research are presented. The third Section will present the CamemBERT model as well as the fine-tuned version that we derived from it in order to classify the productions of L2 French learners. Section IV will be devoted to a discussion of the results.

II. DATASET

The dataset used to fine-tune the model comes from a previous research project on new technologies and their quantitative and qualitative effects on the production of French L2 OCs. The corpus in question is described in [2]. The transcription of this corpus was used as a basis to isolate both the OC and a relevant context of its use. Because of the interview-like nature of this corpus, this resulted in pairs containing a question and the answer to it, constructed as follows:

(3) What have I done/ am I doing with X? You are/were Y it.

Where X is an object, Y is a French verb and *it* is the OC (whose position is mostly preverbal in French). In (4), we have an example where the OC produced by the learner is correct (label 1 in my dataset):

(4) *Qu'est-ce que j'ai fait avec mes crayons? Tu les as rangés*.
English: What did I do with my pencils? You put them away.

In (5), we have an example where the learner uses the repetition of the noun phrase (NP) to avoid using the OC (label 2 in the dataset):

(5) *Que fait la fille avec cette pomme? La fille épluche la pomme*.
English: What is the girl doing with this apple? The girl is peeling the apple.

Finally, in (6), we have several examples of errors in the selection of the OC or in its placement in relation to the verb (label 0):

(6)

a. A misplaced OC: *Qu'est-ce que j'ai fait avec mon crayon? *Tu as l'aiguisé*.
English: What did I do with my pencil? *You have sharpened it.

b. Gender error: *Que font les enfants avec la salade? *Ils le mangent.*
English: What are the children doing with the salad? *They eat it.

c. Number error: *Qu'est-ce que j'ai fait avec les lunettes? *Tu l'as pris dans tes mains*.
English: What did I do with the glasses? *You took it in your hands.

d. d. Grammatical case error: *Que fait le père avec ses enfants? *Il les donne un câlin*.
English: What does the father do with his children? *He gives they a hug.

e. e. Object omission: *Que fait la mère avec son bébé? *Elle regarde*.
English: What is the mother doing with her baby? *She is watching.

As is frequently the case when working with interlanguage, as highlighted by [4], the majority of the examples containing OCs errors are riddled with other errors (lexical spelling, grammatical spelling, or others). As a result, some pairs are labeled 1 despite the fact that there are other errors in the answer, and others are labeled 2 (for repetition) even though the repeated NP is misspelled (e.g., *carte* spelled *cart* or even *card*). We will see that this will not prevent the fine-tuned CamemBERT model from making correct predictions on the submitted data.

The resulting dataset contains 899 question/answer pairs annotated in three categories: 0 for pairs containing errors on OCs, 1 for pairs where the use of OCs is correct, and 2 for pairs where there was a repetition of the NP in the answer. Tab. 1 summarizes the dataset statistics:

TABLE I. DATASET STATISTICS

	N	%
0	126	14.01
1	336	37.37
2	437	48.60

Thus, the distribution of the three categories is unbalanced, as shown in Figure 1.

And since we are dealing with unbalanced categories, the Weighted Random Sampler from the Pytorch library was used to give the less represented data a weight based on their size.

To fine-tune the deep learning model, 80% of the dataset was used for training and the remaining 20% for validation. The next section will be devoted to the description of the model.

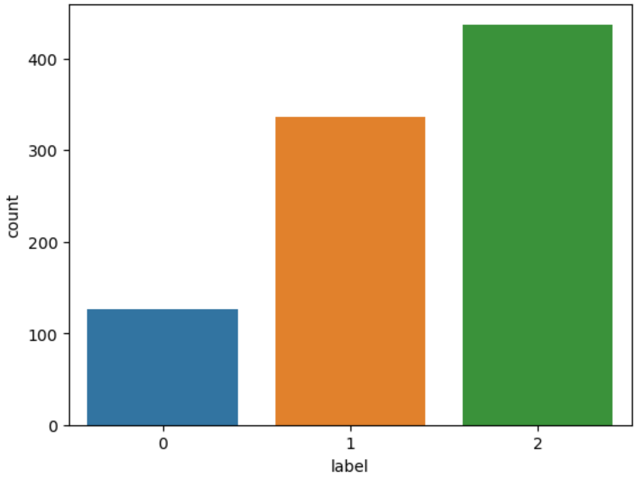

Figure 1. Unbalanced categories in the dataset.

III. CAMEMBERT-BASED MODEL

In this section, the deep learning model that will be used to process object clitics in French is presented. The base model, CamemBERT, will be introduced, as well as the fine-tuned version and its evaluation.

A. *CamemBERT*

Transformer-based [5], CamemBERT, described in [6], is a deep-learning language model for French, based on Bidirectional Encoder Representations from Transformers (BERT), see [7], and more specifically on RoBERTa [8], which "removes BERT's next-sentence pretraining objective, and trains with much larger mini-batches and learning rates". CamemBERT has 110 million parameters and was pretrained on the French subcorpus of the multilingual corpus OSCAR (138 GB of text) as part of a collaboration between INRIA Paris (ALMANACH team) and Facebook/Meta AI.

CamemBERT is suitable for a wide range of NLP tasks, such NER, POS tagging, dependency parsing and natural language inference. Sentiment analysis (see, for instance [9]) through CamemBERT For Sequence Classification python class is another suitable task that led to other applications, such grammaticality judgements. Thus, a fine-tuned version of CamemBERT has been used for coordination error detection in French in [10]. The study by Cheng et al. [11], among many others, used a fine-tuned version of BERT to check the grammaticality of Chinese sentences. Therefore, the goal of this paper is to propose a deep learning model capable of making grammatical judgments by classifying submitted sequences as correct, error-prone, or repetition-prone in order to help French second language learners in better mastering OCs.

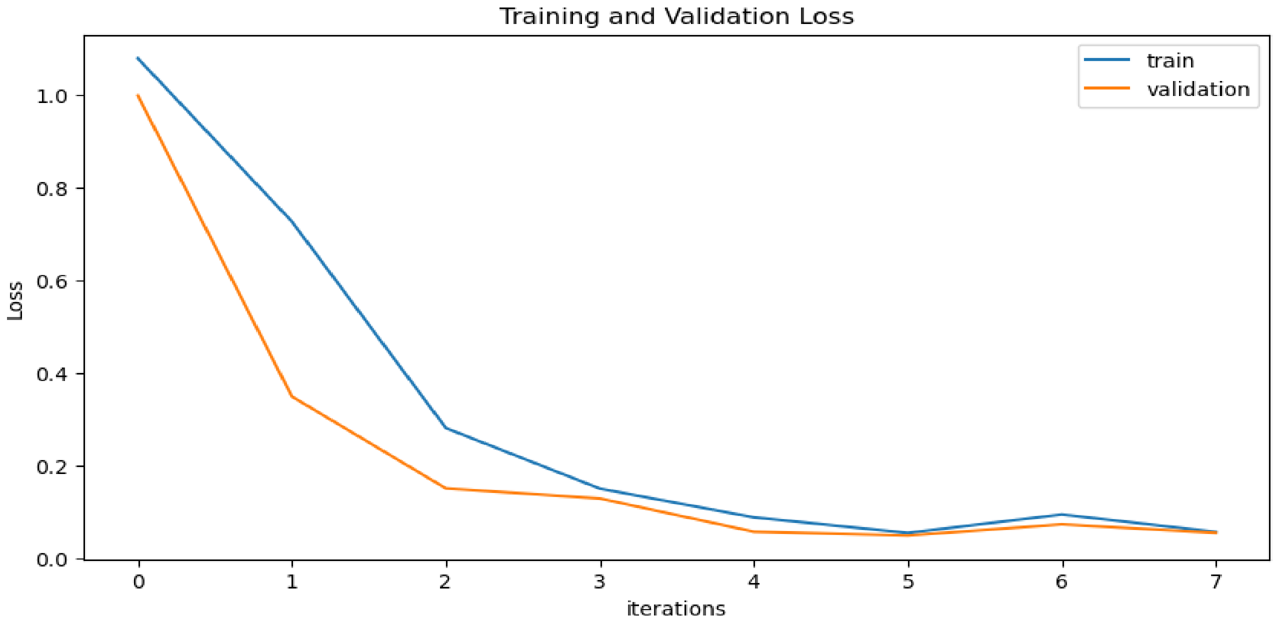

Figure 2. Training and Validation Loss.

B. *The fine-tuned model and its evaluation*

Using the dataset described in Section II, CamemBERT was trained for 10 epochs on an Nvidia consumer GPU with AdamW as the optimizer. The early stop technique was used, which stopped the training at epoch 7, with the Training and Validation Loss that can be seen in Figure 2.

The fine-tuned model obtained an f-score of 0.99. The classification report is shown in Tab. 2.

TABLE II. CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	0.98	0.99	126
1	0.99	1.00	1.00	336
2	1.00	1.00	1.00	437

Figure 3 shows the Confusion Matrix.

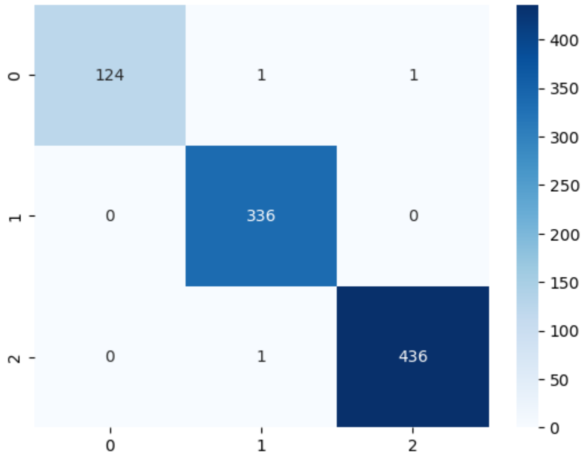

Figure 3. Confusion Matrix.

The model, thus, performed very well on the validation data, with only three pairs misclassified:

(7) *Qu'est-ce que j'ai fait avec le livre? *Vous avez le consulté le livre*.
English: *What did I do with the book? You consulted it the book.
➔ Was classified as 2 (repetition) by the model, but it was labelled 0 (error) in the dataset.

(8) *Que fait l'enfant avec la balle? *L'enfant lance la à son père*.
English: What is the child doing with the ball? *The child it throws to his father.
➔ Was classified as 1 (correct) by the model, but it was labelled 0 (error) in the dataset.

(9) *Que fait la mère avec son bébé? Elle berce son bébé en le regardant*.
English: What is the mother doing with her baby? She rocks her baby while watching him.
➔ Was classified as 1 (correct) by the model, but it was labelled 2 (repetition) in the dataset.

These data will be discussed in Section IV.

IV. DISCUSSION

The model trained on our dataset was able to predict the appropriate grammatical judgment for pairs that it had never seen before but appeared similar to the training data. Here are some examples:

(10) *Qu'est-ce que j'ai fait avec l'échelle? *Tu as la cassée*.
English: What did I do with the ladder? You broke it.
Correct prediction: 0.

(11) *Qu'est-ce que j'ai fait avec le téléphone? Tu l'as donné*.
English: What did I do with the phone? You gave it away.
Correct prediction: 1.

(12) *Qu'est-ce que j'ai fait avec la carte? Tu as sorti la carte*.
English: What did I do with the card? You took the card out.
Correct prediction: 2.

It was also able to correctly make predictions in different contexts of OCs usage. Here are some examples:

(13) *J'ai rencontré Marie et *j'ai lui dit mon secret*.
English: I met Marie and I told her my secret.
Correct prediction: 0.

(14) **Mes amies, je ne peux que l'aimer*.
English: My friends, *I can only love it. Correct prediction: 0.

(15) *Est-ce que les étudiantes aiment la soupe? *Oui, elles aiment.*
English: Do the students like the soup? *Yes, they like.
Correct prediction: 0.

(16) **Je lui aide*.
English: I help him/her.
Correct prediction: 0.

(17) *Je l'observe depuis ce matin*.
English: I have been observing him/her since this morning.
Correct prediction: 1.

(18) *Que mange Marie? *Marie mange*.
English: What is Mary eating? *Mary eats.
Correct prediction: 0.

(19) *Combien de personnes vois-tu? J'en vois trois*.
English: How many people do you see? I see three.
Correct prediction: 1.

(20) *As-tu vu Isabelle? Oui, *je le vois.*
English: Did you see Isabelle? Yes, *I see him.
Correct prediction: 0.

(21) **Elle a le vu*.
English: She saw it/him.
Correct prediction: 0.

The misclassified three pairs (7), (8) and (9) need some explanation. The pair (7), for instance, contains both an error (in the position of the OC) and a repetition of the NP. In this case, the error is more significant and should normally be reported to the user, which was done in the dataset. Pair (8) illustrates the case where the OC is inserted in a postverbal position when it should be in a preverbal position. This error is under-represented in the dataset used for training and accounts for the incorrect label predicted by the model. Finally, pair (9) contains a repetition of the NP and a well-used OC to refer to this phrase. As this repetition here is less troublesome than when the OC is absent, these examples should be annotated differently (as 1) in a future version of the dataset. Therefore, the plan is to conduct a second phase of this research with a completer and more balanced dataset. Regarding the first point, more errors should be represented, such as the one where the OC is postverbal or where the OC is replaced by a strong or tonic pronoun. And in terms of balance, the plan is to create a dataset in which all three classes are equally represented.

V. CONCLUSION

This paper presented a two-part project: the first one consists of setting up an authentic corpus of written productions of learners of French as a second language regarding their use of OCs. This part aims to provide a dataset from the interlanguage in order to carry out the second part. The latter consists of fine-tuning a deep learning model capable of detecting the most frequent errors, but also repetitions and correct sentences.

The main novelty of this approach is to set up a corpus representing interlanguage, which is a glaring lack in research of this type. Moreover, to our knowledge, there have been no previous research projects that aimed to propose a deep learning model to process this interlanguage.

The deep learning model fine-tuned on the described dataset and derived from CamemBERT is robust enough to correctly predict whether a sentence containing an OC in French is correct, incorrect or contains a repetition. This allowed us to develop a graphical interface in Python and PyQT6 to interact with this model. For the time being, this interface only provides a label and a recommendation, but in a future version, we plan to improve the predictive capabilities of the model as well as the feedback refined by error type.

This study, even though it is based on a corpus of only 899 sequences, demonstrates that it is possible to obtain a reliable classifier without resorting to large-scale data, as is the case for current trends in deep learning and generative artificial intelligence, such as the GPT system. In addition, the resulting tool can be an important ally for learners of French as a second language in their quest to master object clitic pronouns, which is one of the difficult aspects of French for these learners. A didactic study is also planned to establish the pedagogical integration of this digital tool for learners at the B1 and B2 levels.

ACKNOWLEDGMENT

This research was made possible by a grant from SSHRC, through the Insight Development Grants Program (2021-2023).

REFERENCES

[1] V. Wust. “The dictogloss as a measure of the comprehension of y and en by L2 learners of French”, The Canadian Modern Language Review, Volume 65, Number 3, pp. pp. 471-499, March 2009.

[2] A. Jebali, “Language anxiety, technology-mediated communication, and elicitation of object clitics in L2 French.”, Alsic 21, 2018. Online. URL : http://journals.openedition.org/alsic/3164, DOI : https://doi.org/10.4000/alsic.3164

[3] L. Emirkanian, L. Redmond, and A. Jebali, “Mastery of dative clitics in ditransitive structures in L2 French by English-speaking learners: influence of argument structure in L1.”, Canadian Journal of Applied Linguistics, Volume 24, Number 3, pp. 30-60, autumn 2021.

[4] A. Affes, I. Biskri, and A. Jebali, “French Object Clitics in the Interlanguage: A Linguistic Description and a Formal Analysis in the ACCG Framework”, in N. T. Nguyen et al. (Eds.): ICCCI 2022, LNAI 13501, pp. 220–231, 2022.

[5] A. Vaswani et al., “Attention Is All You Need”, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, december 2017.

[6] L. Martin et al., “CamemBERT: a Tasty French Language Model”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219, July 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, june 2019.

[8] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, ArXiv, abs/1907.11692, 2019.

[9] G. Guerdoux, T. Tiffet, and C. Bousquet, “Inference Time of a CamemBERT Deep Learning Model for Sentiment Analysis of COVID Vaccines on Twitter”, in J. Mantas et al. (Eds.): Advances in Informatics, Management and Technology in Healthcare, pp. 269-270, 2022.

[10] L. Noreskal, I. Eshkol-Taravella, and M. Desmets, “Erroneous Coordinated Sentences Detection in French Students’ Writings”, in K. Wojtkiewicz et al. (Eds.): ICCCI 2021, CCIS 1463, pp. 586–596, 2021.

[11] Y. Cheng and M. Duan, “Chinese Grammatical Error Detection Based on BERT Model”, Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 108–113, December 2020.