Digitizing 18th-Century French Literature: Comparing transcription methods for a critical edition text

Ann Irvine
Computer Science Dept.
Johns Hopkins University
Baltimore, MD
anni@jhu.edu

Laure Marcellesi
French and Italian Dept.
Dartmouth College
Hanover, NH
laure.marcellesi@dartmouth.edu

Afra Zomorodian
The D. E. Shaw Group
New York, NY

Abstract

We compare four methods for transcribing early printed texts. Our comparison is through a case-study of digitizing an eighteenth-century French novel for a new critical edition: the 1784 *Lettres taïtiennes* by Joséphine de Monbart. We provide a detailed error analysis of transcription by optical character recognition (OCR), non-expert humans, and expert humans and weigh each technique based on accuracy, speed, cost and the need for scholarly overhead. Our findings are relevant to 18th-century French scholars as well as the entire community of scholars working to preserve, present, and revitalize interest in literature published before the digital age.

1 Introduction

Preparing a text for modern publication involves the following: (1) digitizing[1] a printed version of the text, and (2) supplementing the original content with new scholarly contributions such as a critical introduction, annotations, and a thorough bibliography. The second task requires a high degree of expertise and academic insight and the first does not. However, scholars working on such projects often spend large amounts of time transcribing literature from scratch, instead of focusing on skilled contributions.

In this paper, we present an analysis of our efforts using *alternative methods*, other than highly skilled scholars themselves, to transcribe a scanned image of a novel into a modifiable, searchable document. We compare four different methods of transcription with a gold standard and evaluate each for accuracy, speed, and cost-effectiveness. Choosing an appropriate transcription method may save scholars time and allow them to focus on critical contributions.

First published in 1784, Joséphine de Monbart's *Lettres taïtiennes* is an epistolary novel dramatizing the European colonial takeover of the newly-encountered island of Tahiti from the fictional point of view of a young Tahitian woman. While most works of the time painted a fictional Tahitian paradise of uninhibited sexuality, this novel offers a singular anti-colonial critique by grounding it in the suffering of the female body. We describe our work transcribing the second edition of the novel, which is written in French and was published in Paris, without date (probably 1786). The text is comprised of 156 pages, which are split into two volumes.

There are many off-the-shelf (OTS) natural language processing (NLP) tools available for French, including optical character recognition (OCR), context-sensitive spell checking, and machine translation. Additionally, French is a widely spoken language in the world today and it is often possible to recruit French speakers to do transcription and annotation. However, the early-modern (18th-century) form of the language varies substantially from the modern form, which is used to train OTS French tools and is what non-domain-expert transcribers are familiar with. Differences between the modern and early-modern forms of the language include orthography, lexical choice, and morphological patterns.

An additional challenge is that our transcriptions are based on a *copied* version of the bound text available at the Bibliothèque nationale de France. This common scenario introduces the challenge of noise, or ink marks which are not part of the text. Scattered dots of ink may result in punctuation and character accenting errors, for example.

In this paper, we compare the accuracy, speed, and

[1] In this work, *digitizing* means transcribing an image into a modifiable, searchable file of unicode characters.

Workshop on Computational Linguistics for Literature, pages 64–68,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

cost of using several different methods to transcribe *Lettres tahitiennes*. In Section 2 we describe the transcription methods, and in Section 3 we present a detailed analysis of the types of errors made by each. We also provide a discussion of the difficulty of post-editing the output from each transcriber. Section 4 gives an overview of prior work in the area and Section 5 a practical conclusion, which may inform scholars in the beginning stages of similar projects.

2 Methods

We compare four sources of transcription for 30 pages of the novel with one gold standard:

- OTS French OCR output
- Non-expert French speakers on Amazon's Mechanical Turk (MTurk)
- Non-expert undergraduate students in the humanities, closely supervised by the expert
- Professional transcription service
- Gold standard: early-modern French literature scholar and editor of the critical edition

Given PDF images of a copy of the novel, each source transcribed the same 30 pages[2]. The pages are a representative sample from each of the two volumes of the text.

We used OTS Abbyy Finereader OCR software, which is trained on modern French text and has a fixed cost of $99.

Three MTurk workers transcribed each page of text, and the domain expert chose the best transcription of each page. In future work, we could have another round of MTurk workers choose the best transcription among several MTurk outputs, which has been shown to be effective in other NLP tasks (Zaidan and Callison-Burch, 2011). We paid each MTurk worker $0.10 to transcribe a single page.

Two closely supervised undergraduate students transcribed the novel[3], including the 30 test pages. The cost per page per student was about $0.83.

Our group also hired a professional company to transcribe the entire novel, which charged a fixed cost of $1000, or about $3.21 per page.

The early-modern French literature domain-expert also transcribed the 30 test pages from scratch, and this transcription was used as the gold standard for measuring accuracy.

Because the critical edition text should be as faithful as possible to the original text, with no alteration to the spelling, syntax, capitalization, italicization, and paragraph indentation, we define as errors to be:

- an incomplete transcription
- missing or added words, letters, or characters
- a word transcribed incorrectly
- capitalization, bold, italics not matching the original text
- incorrect formatting, including missing or added paragraph indentations and footnote distinctions

In Section 3, we present a quantitative and qualitative analysis of the types of errors made by each of our transcription methods.

3 Results and Error Analysis

Table 1 lists the error rate for each transcriber.

3.1 S/F errors

One of the most common errors made by all four transcription methods is confusing the letter **ſ** (or long s), which is common in early-modern French but doesn't appear in modern French, with the letter **f**. Figure 1 shows examples of phrases in the original document that include both characters. These examples illustrate how familiarity with the language may impact when transcription errors are made. All three human transcribers (MTurk workers, students, professionals) confused an **f** for an **ſ** in (b). Interestingly, the phrase in (b) would never be used in modern French, so the transcribers, not recognizing the overall meaning of the sentence and wary of 'missing' a ſ, incorrectly wrote *seront* instead of *feront*. In contrast, the phrase in (a) is rare but does exist in modern French. The MTurk worker and professional transcriber correctly transcribed *feront* but the student, who probably didn't know the phrase, transcribed the word as *seront*.

The OCR system trained on modern French did not recognize **ſ** at all. In most cases, it transcribed the letter as an **f**, but it sometimes chose other letters, such as **t**, **i**, or **v**, in order to output French words that exist in its dictionary. Although it may have been

[2] Each page is in the original duodecimo format and contains about 150 word tokens.

[3] One student transcribed volume 1, the other volume 2.

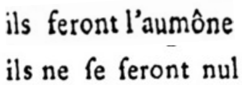
ils feront l'aumône
ils ne ſe feront nul ſcrupule

Figure 1: Correct transcription: (a) ils feront l'aumône (*give alms*). The student incorrectly transcribed *feront* as *seront*. (b) ils ne se feront nul scrupule (*they will have no qualms*). All four alternative transcription sources incorrectly transcribed *feront* as *seront*.

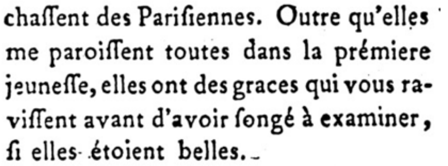
chaſſent des Pariſiennes. Outre qu'elles
me paroiſſent toutes dans la prémiere
jeuneſſe, elles ont des graces qui vous ra-
viſſent avant d'avoir ſongé à examiner,
ſi elles étoient belles.

Figure 2: Correct transcription: Outre qu'elles me paroissent toutes dans la prémiere jeunesse, elles ont des graces qui vous ravissent avant d'avoir songé à examiner, si elles étoient belles (*Besides [these women] appearing to me in the prime of youth, they have graces that delight you before you even think of considering whether they are beautiful*. Transcribers made both conjugation (*paraissent* vs. *paroissent*) and accenting (*prémiere* vs. *première*) modernization errors in this passage.

possible to train the OCR system on early-modern French, the very slight difference between the character strokes means that disambiguating between **f** and ſ would likely remain a difficult task.

3.2 Modernization errors

Eighteenth-century French is understandable by speakers of modern French, but there are a few differences. In addition to the absence of the letter ſ, modern French conjugates verbs with $-ai, -ais, -ait, -aient$ instead of $-oi, -ois, -oit, -oient$ and follows stricter rules that no longer allow for variations in spelling or accenting. Figure 2 shows examples of both. In general, the authors of modern critical editions seek to maintain original spellings so that future scholars can work as close to the original text as possible, even if the original work includes typos, which we have seen. However, our human transcribers incorrectly modernized and 'fixed' many original spellings. This is likely due to the fact that it is hard for a human transcriber who is familiar with the language to *not* 'correct' a word into its modern form. We observed this across all human transcribers. For example, our professional transcriber transcribed *première* instead of *prémiere* and one MTurk worker transcribed *chez* instead of *chés*. The OCR model, which is trained on modern French, is also biased toward modern spellings. Most of its modernization errors were related to accents. For example, it transcribed *graces* as *grâces* and *differentes* as *différentes*.

Some modernization errors occur systematically and, thus, are easy to automatically correct after the initial transcription is complete. For example, all $-aient$ word endings could be changed to $-oient$. This is not true for all modernization errors.

3.3 Errors from page noise

Since all of our transcribers worked from a scan of a copy of the original book held at the Bibliothèque nationale de France, noise in the form of small dots, originally bits of ink, appears on the pages. These small dots are easily confused with diacritics and punctuation. Our human transcribers made such errors very infrequently. However, this type of noise greatly affected the output of the OCR system. In addition to mistaking this type of noise for punctuation, sometimes it affected the recognition of words. In once instance, *visages* became *ylfygc* because of small dots that appeared below the v and a[4].

3.4 Formatting errors

We asked all transcribers to maintain the original formatting of the text, including paragraph indentations, footnotes, and font styles. However, because of limitations inherent to the MTurk task design interface, we were unable to collect anything but plain, unformatted text from those transcribers. In general, our other human transcribers were able to accurately maintain the format of the original text. The OCR output also made formatting mistakes, particularly bold and italicized words.

3.5 Other errors

Both humans and the OCR system made an assortment of additional errors. For example, two MTurk workers failed to turn off the *English* automatic spell correctors in their text editors, which resulted in *lettre* becoming *letter* and *dont* becoming *don't*.

3.6 Scholar overhead

Table 1 lists the average number of errors per page for each transcription method. In addition to consid-

[4] In this example, an ſ was also transcribed as an **f**

Error	OCR	MTurk	Prof.	Stud.
Modernization	26.29	2.82	0.71	0.46
Noise	7.68	0.0	0.32	0.21
Formatting	1.96	0.82	0.36	0.0
Total	35.93	3.86	1.39	0.71

Table 1: Mean number of errors per page, by error type and transcription method. The total includes the error types shown as well as an assortment of other errors.

ering the error rate of each, we found that it is critical to consider (a) the effort that the scholar must exert to correct, or post-edit, a non-expert's transcription, and (b) the amount of overhead required by the scholar to gather the transcriptions.

All errors are not equally serious. We found that the expert scholar had an easier time correcting some errors in post-editing than others. For example, modernization errors may be corrected automatically or in a single read through the transcription, without constantly consulting the original text. In contrast, correcting formatting errors is very time consuming. Similarly, correcting errors resulting from page noise requires the scholar to closely compare punctuation in the original text with that of the transcription and takes a lot of time.

Previous research on gathering and using non-expert annotations using MTurk (Snow et al., 2008; Callison-Burch and Dredze, 2010; Zaidan and Callison-Burch, 2011) has been optimistic. However, that work has failed to account for the time and effort required to compose, post, monitor, approve, and parse MTurk HITs (human intelligence tasks). In our exploration, we found that the time required by our expert scholar to gather MTurk annotations nearly offsets the cost savings that result from using it instead of local student or professional transcribers. Similarly, the scholar had to provide some supervision to the student transcribers. The professional transcription service, in contrast, though more expensive than the other high quality (non-OCR) methods, required no oversight on the part of the scholar. After using all methods to transcribe 30 pages of *Lettres taïtiennes* and critically comparing the costs and benefits of each, we had the professional transcription service complete the project and our expert French literature scholar has based a new critical edition of the text on this transcription.

4 Background

Snow et al. (2008) gathered annotations on MTurk in order to supervise a variety of NLP tasks. In general, they found a high degree of annotator agreement and inspired a plethora of research on using non-expert annotations for additional tasks in language processing (Callison-Burch and Dredze, 2010).

OCR has been an active area of research in NLP for decades (Arica and Yarman-Vural, 2001). Recent work has acknowledged that post-editing OCR output is an important engineering task but generally assumes large amounts of training data and does not attempt to maintain text format (Kolak et al., 2003). As we described, for our application, transcribing all content and formatting, including footnotes, references, indentations, capitalization, etc. is crucial. Furthermore, OCR output quality was so low that post-editing it would have required more work than transcribing from scratch. We did not attempt to train the OCR since, even if it had recognized **ſ** and learned an appropriate language model, the formatting and noise errors would have remained.

5 Future Work and Conclusions

In Section 3.2, we mentioned that it may be possible to automatically post-edit transcriptions and correct systematic modernization errors. The same may be true for, for example, some types of typos. This type of post-editing could be done manually or automatically. One potential automatic approach is to train a language model on the first transcription attempt and then use the model to identify unlikely segments of text. We plan to pursue this in future work.

Although we hoped that using MTurk or OCR would provide an inexpensive, high-quality first round transcription, we found that we preferred to use student and professional transcribers.The tradeoffs between speed and accuracy and between low cost and overhead time were not worthwhile for our project. If a scholar were working with a larger text or tighter budget, investing the time and effort to use MTurk could prove worthwhile. Using an OCR system would demand extensive training to the text domain as well as post-editing. This paper enumerates important challenges, costs, and benefits of several transcription approaches, which are worthy of consideration by scholars working on similar projects.

References

N. Arica and F. T. Yarman-Vural. 2001. An overview of character recognition focused on off-line handwriting. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 31(2):216–233, May.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June. Association for Computational Linguistics.

Joséphine de Monbart. 1786. *Lettres tahitiennes*. Les Marchands de nouveautés, Paris.

Okan Kolak, William Byrne, and Philip Resnik. 2003. A generative probabilistic ocr model for nlp applications. In *Proceedings of the NAACL*, pages 55–62. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from nonprofessionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA, June. Association for Computational Linguistics.