Journal of French Language Studies (2022), **32**, 216–242
doi:10.1017/S0959269522000011

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

The variable use of diacritics in synchronous computer-mediated French discourse: Replication research

Geraldine Blattner[1] and Lawrence Williams[2,*]

[1]Florida Atlantic University, LLCL, Florida, United States and [2]University of North Texas, World Languages, Literatures & Cultures, Texas, United States
*Corresponding Author. Email: lawrence.williams@unt.edu.

(Received 09 March 2021; revised 18 December 2021; accepted 04 January 2022)

Abstract
This study promotes replication research as a methodological approach that is needed in order to compare earlier and more recent analyses of digital discourse. When much of the existing research was conducted, the primary means of communication included the use of a computer keyboard, (presumably) less bandwidth, and fewer devices. However, with an increase of the range of device types, the study of diacritics deserves another look within the Digital Media landscape. The present study examines the variable use of diacritics in synchronous (i.e., real-time) French chat discourse. We have replicated a study with different data sets from the same chat corpus, which is composed of data from a European chat server. We have also compared the data from the 2008 half of the corpus to data from the same chat channels collected in 2016 (just over 60,000 words in each half of the corpus, which included a total of 7,569 tokens that were coded). Our analysis of the 2008 corpus showed that one main finding was not the same as ours (from a different part of the 2008 corpus). Moreover, a diachronic analysis (2008 vs. 2016) revealed reversed trends between the two age-based channels (i.e., 20s vs. 50s).

1. INTRODUCTION

Computer-mediated communication has become commonplace in our digital world; consequently, electronic discourse has been the focus of an ever-growing number of linguistic investigations, some of which have focused on orthographic variation (Anis, 1999; Pierozak, 2000; Crystal, 2001; Myslín and Gries, 2010; van Compernolle and Williams, 2010; Yus, 2010; among others). Discourse in some computer-mediated communication environments (e.g., synchronous chat, mobile texting) resembles traditional oral genres as they frequently take on characteristics of playful and stylized performance (Bauman, 1992; Finnegan, 1992; Anis, 1999; Pierozak, 2003; Sykes, 2005). Of course, some types of computer-mediated discourse (e.g., e-mail, discussion boards) resemble traditional written discourse (Yates, 1996; Williams, 2009). Consequently, for

© The Author(s), 2022. Published by Cambridge University Press.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

either traditional or electronic discourse, it is often not possible to provide absolute, categorical characterizations and descriptions since several social and technical factors can determine or influence both the process and the product (Herring, 2007). Therefore, the present article focuses on the exploration of age as a possible factor that influences language variation in one type of computer-mediated discourse.

Few studies have focused on language variation in computer-mediated French discourse (van Compernolle and Williams, 2007, 2010), including the presence or the omission of the five diacritics used in French: the acute (é), grave (à, è, ù), and circumflex (â, ê, î, ô, û) accents, as well as the cedilla (ç) and the diaeresis (ë, ï, ü) in electronic discourse, which is the focus of this study. One goal of the present study is to provide a more complete representation of what is happening in terms of the use of diacritics in this digital type of discourse and to establish the extent to which the informal nature often associated with such discourse results in a systematic orthographic modification of words containing diacritics. We explore the distribution of the diacritics used in French according to their linguistic environment (the letter they accompany), the (self-reported) age of the participants, and the period of the interaction. Given that relatively little previous work has examined the variable use of diacritics, this study seeks to expand our understanding of the variable use of diacritics in synchronous computer-mediated French discourse by replicating a previous study.

We have adopted a variationist framework for our analysis of diacritics because we are interested in the same underlying question that originally inspired Labov (1963, 1972) many years ago: How do people say the same thing in different ways? A review of variationist sociolinguistics is beyond the scope of the present study; however, we encourage readers who are interested in exploring the framework inspired by Labov to read the comprehensive treatment by Tagliamonte (2012).

The remainder of this article is organized in the following way:

- Overview of French orthography and diacritics;
- Computer-mediated French discourse;
- Study 1, Blattner and Williams 2008 corpus (replication of Study 0 by Al-Rashdan, 2010);
- Study 2, Blattner and Williams 2016 corpus (replication of Study 1); and
- Conclusion.

Since this study is not organized as a traditional research project using a single corpus, we have provided a summary (see Table 1) of basic information about the data sets and the corpora that we mention throughout this article.

We refer to the study by Al-Rashdan (2010) as *Study 0* (i.e., zero) because we are only using this as a point of comparison with Study 1; we used different data sets from the same corpus in order to determine if a replication (BW08 corpus) would produce similar results. Study 2 is our analysis of data that we collected almost a decade later (BW16 corpus) from the same chat channels.

Since the focus of the present article is to provide replication research, we have two research questions designed to guide this type of study:

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 1. Corpora

Study	Corpus	Label	Words
0	2008	Al-Rashdan (2010)	59,977
1	2008	BW08 corpus	60,364
2	2016	BW16 corpus	60,020

1) To what extent – regarding the variable use of diacritics – do the results of a replication study resemble the initial results from a different subset of data within the same corpus?
2) To what extent – regarding the variable use of diacritics – do the results of a replication study resemble the initial results from data in a corpus collected almost ten years earlier from the same source?

2. FRENCH ORTHOGRAPHY AND DIACRITICS

In this section, we offer an overview of diacritics used in French, especially for readers who are not familiar with the variable use of these symbols in traditional or digital French discourse. We recognize that scholars in the fields of French Cultural Studies or French Linguistics will already be very familiar with the use of diacritics and the spelling reforms that have been undertaken in France and elsewhere in the francophone world during the twentieth century.

The appearance and standardization of diacritics in the French language is the focus of one of the many interesting debates that has unfolded over the past few centuries, but it is also a contemporary issue that is regularly revisited.

Accents and other diacritics first appeared in the sixteenth century and were progressively accepted and widely used in the orthography of French. Rickard (1993) explains that the acute accent in word-final masculine [e] was the first diacritic extensively used by printers. In the seventeenth and eighteenth centuries, the standardization of all diacritics took place under the watchful collective eye of the *Académie française* (Pellat and Andrieux-Reix, 2006) and was still an ongoing process in the nineteenth and twentieth centuries as numerous reform proposals were developed (Watts, 1991; see also Al-Rashdan, 2010). In 1989, Michel Rocard, the Prime Minister of France at the time, created *le Conseil supérieur de la langue francaise*, which, one year later, proposed *des aménagements orthographiques* designed to eliminate numerous orthographic anomalies and *absurdités* in the French language. Some of the 2,000 orthographic *rectifications* published in the *Journal officiel de la République francaise* on 6 December 1990 focused on diacritics. For instance, the use of the circumflex accent on the letters *i* and *u* for the spelling of words like *île* 'island', *bûche* 'log', and *soûlard* 'drunkard' was deemed to be no longer necessary. In words like these, there is no underlying motivation for the use of the circumflex accent since it does not indicate a specific pronunciation or distinguish two homographs such as *jeune* 'young' and *jeûne* 'fasting'. It is important to note, for

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

example, that these rectifications do not apply to proper nouns, such as *Benoît 'Benedict'* and N*îmes*, a city in southern France.

Additional modifications were made to words containing an acute or grave accent. In French, the acute accent can accompany the vowel *e*, indicating the pronunciation of a closed [e] like in the word *bébé*. On the other hand, the grave accent can be placed on several vowels (*e, a, u*); on the vowel *e* it indicates that it should be pronounced [ɛ] as in the word *père*. A series of simple rules regulate the usage of these accents.

The acute accent appears on *e* at the beginning of a word; it often indicates that an *s* used to follow that vowel, as in the word *étudiant*. The same accent can appear on a syllable-final *e*, as in *blé*, *Fé-lix*. In line with these rules, the *Conseil supérieur* added acute accents on words which contain the pronunciation of a closed [e] yet were listed in the *Académie française* dictionary without any accent (i.e. *québécois, réfréner, réclusionnaire*). In addition, as part of the same orthographic reform, words of latin or foreign origin (e.g., facsimile, sombrero) are recommended to be written with an acute accent – facsimilé, sombréro – in order to reflect their French pronunciation (Catach, 1991: 37).

The vowel *e* takes a grave accent when it follows a vowel or consonant and is followed by a syllable containing a mute *e*, commonly called schwa as in *fidèle, règlement*. As a desire to make this rule systematic, the *Conseil supérieur* modified the spelling of a limited number of words which used to be written with an acute accent and now take a grave accent. Such words are *abrègement, assèchement, crèmerie, évènement* and *fèverole*. Additional modifications were adopted regarding the use of this accent in verbal inflections; however, in the present article, we do not provide the entire list of l*es aménagements orthographiques* adopted by the French government. This section only serves as an introduction to some of the issues surrounding French orthography. In any case, we wish to note that the spelling reform of 1990 in France has been more or less ignored by many teachers and dictionary editors/publishers (see Daussy, 2009, for an overview).

It is important to note that the present study focuses on the variable use of diacritics, and this implies that we have not undertaken a more global review of orthographic variation or any kind of system for categorizing different forms. Nonetheless, we recognize that one aspect of most studies that explore computer-mediated communication involves the debate over how to classify different types of communication, whether they be older or more recent. Work by Gadet (2008), for example, has highlighted the complexities of orthographic variation in different types of computer-mediated communication in order to explore the notion of oral-written hybridity (see also Anis, 1999; Pierozak, 2000). This is in line with Werry's (1996) analysis of synchronous chat, which he labels as *interactive written discourse*.

> From conveying patterns of intonation, to supplying paralinguistic cues, to the collaborative construction of a shared contextual frame, language on IRC is crucially inflected by the practice of naturalistic interpersonal exchange, and the tendency to create forms of expression that enable or are appropriate to such exchange. (61)

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Gadet (2008: 523) seems to view computer-mediated communication not as a genre, but rather as a context for communication in which users adapt their style in order to be efficient, humorous, or creative; adapting one's style can also be done in order to create a counterculture or countercultural elements within a given communication space.

Stark (2015) also explores the complexities of oral and written influences that appear in digital discourse, but with a specific focus on mobile text messages. In her corpus of Swiss text messages (collected in 2011) written in French, Stark found "very high percentages of standard orthographic marking of morphosyntactic information … on the one hand, and interesting deviations from the standard norm … " (395). This demonstrates – once again – that there is no absolute way to categorize computer-mediated communication as more oral or more written because there are so many different factors that can influence language use. Readers interested in different theoretical approaches to studying French orthography would greatly benefit from consulting the foundational work by Catach (1973, 1980), Catach, Gruaz and Duprez (1986), and Anis (1988), as well as a more recent overview of French orthography by Fayol and Jaffré (2016).

3. COMPUTER-MEDIATED DISCOURSE

Research that has explored various types of orthographic variation (e.g., Anis, 1999; Pierozak, 2003; van Compernolle and Williams, 2007, 2010) in computer-mediated discourse (CMD) suggests that these phenomena are often primarily motivated by the rapidity of the interactions and the desire for efficiency. As van Compernolle (2011) explains, 'written forms are not always convenient to produce in electronic environments' (134); consequently, Internet users frequently decide to omit accents or use abbreviations as it is a more expedient option if it does not lead to linguistic ambiguity or some other type of misinterpretation, depending on the context in which diacritics could appear. It is therefore not surprising that certain types of orthographic variation are more widespread in synchronous electronic discourse (i.e. chats) than asynchronous environments (i.e. fora, blogs) (van Compernolle and Williams, 2010; Cougnon, 2010).

Over the past two decades, studies in the area of computer-mediated discourse (CMD) have been conducted at an increasing rate, yet only a handful of studies have investigated the use of diacritics in CMD. Research by van Compernolle and Williams (2010) investigated the orthographic variation of the French acute accent in two types of electronic environments. Their statistical analysis revealed that *é* is used much less frequently in synchronous chat than in asynchronous forums. As previously established with other variables (van Compernolle and Williams, 2007), the level of formality of the discourse produced in different CMD environments (e.g., synchronous vs. asynchronous) can vary noticeably. Their analysis demonstrated a correlation between the formality associated with a particular electronic environment and the amount of standard/formal orthographic variables observed within a single computer-mediated context. The same study also identified the position of the *é* as a contributing factor in the variation of use of the acute accent. Results showed that in synchronous chat *é*

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

(i.e., with an acute accent) is particularly disfavoured in word-internal position and more likely to be maintained in word-initial or word-final position. Therefore, van Compernolle and Williams (2007) concluded that even if some diacritics appeared to be used less frequently in some types of electronic discourse than in standard writing, it is clear that they have not disappeared from electronic discourse.

Al-Rashdan (2010) used age as one factor in his investigation of all the possible diacritics associated with the letters *a*, *e*, *i*, *o*, *u*, and *c* in synchronous French-language chat. A corpus-driven apparent-time variationist analysis revealed that participants use diacritics at an overall rate of 55.4%. The letter *e* was the most frequently associated with either the acute, grave or circumflex accents by the chat participants. Al-Rashdan established that – in general terms – the participants favoured the use of diacritics not only on the letter *e*, but also used the cedilla on the letter *c*. Conversely, the results showed that diacritics occurred much less frequently with the letter *i* (5.9%) in the online exchanges analysed. In terms of the specific diacritics, the acute accent appeared to be the most systematically used regardless of the letter it accompanied, whereas the circumflex accent is almost always omitted. Al-Rashdan also conducted an analysis of the use or the omission of diacritics in relation to the age of the participants. The nature of the data allowed for such an analysis, as the participants either joined the *vingtaine* 'people-in-their-twenties' (20s) chat or the *cinquantaine* 'people-in-their-fifties' (50s) group. His results revealed no statistically significant differences between these two age groups (20s and 50s), suggesting that older and younger synchronous chatters demonstrate similar patterns for this type of orthographic variation.

In another recent study that explored age as a factor that can influence language variation, van Compernolle (2011) explored the use of diacritics (i.e., cedilla; acute, grave, and circumflex accents) on Canadian dating sites. The results showed that both age and gender appeared to play a noticeable role in patterns of the use or omission of diacritics in this type of asynchronous electronic discourse. Self-described older users of the sites favoured the use of diacritics and so did women. Furthermore, van Compernolle established that the acute accent was the least omitted diacritic in his data in comparison to the cedilla and the circumflex and grave accents. He argued that the most likely explanation for the presence of diacritics in online written language can be associated with prestige, which also implies and presupposes correctness. He suggested that by maintaining diacritics, older women showed 'the desire to dissociate oneself from a perceived "subordinate" variety of (online) language, characterized by non-standard, abbreviated writing' (145).

Even if Tatossian (2008) did not explore the variable use of diacritics, her analysis is relevant to the present study since it examines French synchronous chat discourse. In a corpus of 4,520 chat messages (2,260 on channels for adults and 2,260 on channels for adolescents), Tatossian found that there were indeed some noticeable differences between the discourse produced by these two groups. At a general level, there was a tendency for adolescents to use more creative spellings in cases where using different letters would produce the same sound (when read on a computer screen). For example, the letters *au* are pronounced – in many contexts – like the letter *o*, so using *bo* to mean *beau* 'handsome'/'pretty'/

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

'beautiful' (or an equivalent substitution) occurred in 65.8% of cases in discourse produced by adolescents and in 52.7% of cases in discourse produced by adults. In Tatossian's overview of the data, the only main category in which adults used more variation than adolescents was the use of linguistic creativity in order to accentuate or emphasize words and feelings (e.g., repeating the letter *l* several times at the end of *lol*, meaning *laughing out loud*, which also happens to be used in French). After presenting four general categories of the analysis, Tatossian provided a detailed analysis of each of these categories in order to demonstrate that even though three of the four general categories suggested that adolescents engaged in more linguistic variation, there are specific aspects of digital discourse that show more variation in the discourse produced by adults. In the category that examined abbreviated forms, for example, there was a tendency for adults to use use acronyms and initialisms (41.8% for adults; 20.0% for adolescents). This was also the case for aphesis (4.0% for adults; 2.6% for adolescents); however, this sub-category had relatively few tokens. Overall, the detailed results provided by Tatossian revealed that even though adolescents and adults used non-traditional forms at different rates, there was consistency for both groups related to the majority of tokens. In other words, if one group used a traditional form either more or less than 50% of the time, the other group did so too, but in some cases the difference was noticeable. From a holistic perspective, Tatossian's results reinforce the notion that it is impossible to make an absolute, across-the-board claim about either younger people or adults producing more or less variation in all contexts, suggesting that age does not necessarily seem to be the determining factor that dictates orthographic choices in digital exchanges.

Although it seems reasonable to suggest that non-standard typography (NST) is used for the sake of rapidity of online interactions, there is indeed research that shows how various other factors can influence different ways in which people shape digital communication. In an exploration of NST in text messages (i.e., SMS) and Facebook messages within the same context (i.e., social media for an Italian radio station) Zelenkauskaite (2017), took into consideration message length and addressee as two variables. Using a corpus of 1,000 text messages and 1,000 Facebook messages over a period of four months in 2011, Zelenkauskaite found that there was "no [statistically significant] difference in message length between Facebook and SMS messages" (131). Other findings were that "Facebook had more overall NST than SMS" (131), "SMS had fewer insertions than Facebook" (131), "SMS had more abbreviations than Facebook" (131), and "messages addressed to one person contained more insertions for Facebook than to the radio [i.e., the radio station as the owner and organizer of the on-air programming and the social media], and more abbreviations in SMS addressed to one person than to the radio" (131). This type of analysis demonstrates that it might be worthwhile in other modes of communication (e.g., Internet Relay Chat) to make more of an effort to distinguish different categories of addressees as a way to expand CMC analysis.

Taking into account various motivational factors can also prove to be worthwhile in any analysis of CMC. For example, Zelenkauskaite and Gonzales (2017) examined a corpus of 1,534 text messages produced by 35 participants during a

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

period of three and a half months in order to explore NST. "All members of each group lived in the same housing development – i.e. in New York City public housing developments. All participants spoke English and 70% also spoke a second language (primarily Spanish)" (78). Zelenkauskaite and Gonzales "found no effect of age or gender on frequency of overall NST or NST by individual category [i.e., NST in words, non-standard punctuation, non-standard use of spaces, and use of emoticons]" (82). In fact, the overall results revealed "that the majority of the messages did not include non-standard typography use (55.7%) and the most frequently used was one type of non-standard typography, with only one message using all four types of non-standard typography" (82). In order to contribute to the debate over NST as "poor literacy skills … or symbolic capital within a group" (83), Zelenkauskaite and Gonzales focused on demographic variables. Their analysis showed that "participants with greater education were more likely to use non-standard typography than participants with less education. There were no other demographic effects on overall non-standard typography use" (83). These findings provide support for "a *symbolic capital hypothesis* of NST use, suggesting that participants use NST to signal status and expertise within their digital community rather than reflect a lack of knowledge or literacy … " (86) Additional support for a symbolic capital hypothesis of NST, according to Zelenkauskaite and Gonzales, is also supported by the participants' increased use of "NST over time, despite an overall drop in texts" (86).

It is also possible to view NST as symbolic capital in the context of a "dating market" (Herring and Zelenkauskaite, 2009: 5). The authors of this study explored CMC using a corpus of 1,452 text messages sent by members of the public to an "Italian interactive television (iTV) program" (5). Herring and Zelenkauskaite focused on trying to determine whether or not there were "gender differences in abbreviation and insertion in iTV SMS and, if so what they are, how frequently they occur, and what they mean" (6). Since gender was a main factor for their study, messages in which no gender indication could be found were excluded, as were some other types of messages, resulting in an adjusted corpus of 1,164 messages for the analysis. The results of this study contradicted two of the authors' hypotheses (based on previous research) and supported their third hypothesis.

> Underlying all of these interpretations is the notion that typographic (non) standardness is a form of social capital in iTV SMS, much as phonological (non)standardness is in speech, and that local markets (e.g., the dating and heterosexual markets proposed here, or the informal social marketplace in recreational CMC more generally) can exist within larger linguistic markets (Bourdieu, 1977), assigning different values to (non)standardness according to the context. (26-27)

As the authors noted, only one iTV program was analysed, and it is quite possible that other programs with other topics might produce noticeably different results.

In another study that explored potential differences between the language produced by men and women in CMC, Zelenkauskaite and Herring (2006)

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

focused on typographical elements in Lithuanian and Croatian synchronous chat. Given the technological constraints of IRC (at the time of their study and perhaps even in some contexts today), Lithuanian and Croatian users were "not able to produce standard spellings of some words in their languages in IRC" (475). Users of English, for example, happen to be able to use English letters in ASCII character systems for text-based CMC; however, this is not the case for users of many other languages, who "are often forced to resort to creative orthographic and typographic 'work-arounds' … [e.g., not only users of Lithuanian and Croatian, but also users of languages such as Arabic and Greek]" (475). The analysis for this study was based on a corpus of synchronous chat data collected in 2001 from IRC channels on a Lithuanian server and a Croatian server. The authors found some noticeable typographical features that revealed noticeable differences between male and female users.

> The results of our study of five IRC chat channels in each language show that both technologically conditioned and voluntary deviant typography are correlated with the gender of the user. In both Lithuanian and Croatian, the characters š, ž, č cannot be typed in IRC. Women tend to use s, z, c, while men prefer the longer forms sh, zh, ch. Another finding is that female users, especially in Lithuanian, voluntarily use "softer," palatalized consonants. (475)

Even though most people might imagine that technology has developed so much over the past couple of decades so that anyone can produce all characters in every language, it seems that there will most likely always be constraints related to producing all characters on all devices in all contexts.

Another aspect of CMC that can influence variation is playfulness. This topic has been explored by Danet, Ruedenberg-Wright, and Rosenbaum-Tamari (1997), among others. In this study, the authors categorized play into three different types: "1) play with identity; 2) play with frame of interaction; and 3) play with typographic symbols" (Abstract). Their analysis of a virtual party in a synchronous chat room revealed that improvisation was quite common and that many aspects of the online party resembled actions and behaviours that typically happen in real life.

4. RESULTS FOR STUDY 1 – BW08 CORPUS; REPLICATION OF AL-RASHDAN (2010)

4.1 The Corpus

The data for this study were taken from two real-time chat channels based on the age groups *vingtaine* (20s) and *cinquantaine* (50s) on the EuropNet website, a public federation of European Internet Relay Chat networks. The corpus analysed below consists of several data sets that form two approximately equal halves of the corpus, whether it is divided by the month of data collection or the (perceived) age group of the participants (see Table 2). We limited our corpus to approximately 60,000 words since this study aims to replicate the study by Al-Rashdan (2010).

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 2. Word count in the BW08 corpus for Study 1

Session	Vingtaine	Cinquantaine	Total
1 (May)	13,433	16,615	30,048
2 (June)	15,877	14,439	30,316
Total	**29,310**	**31,054**	**60,364**

4.2 The participants

As is often the case for studies of CMD, little to nothing is known about the participants. Even the scant amount of personal, demographic information available is often unverifiable. However, when participants belong to a particular age group in a chat room, this known characteristic can be isolated as an independent variable and be taken into consideration as an indication of linguistic behaviour in general or, more specifically, as a measure of the use or the omission of diacritics. Nonetheless, it is important to note that self-reported data in any context is not completely reliable. In the case of synchronous chat, there are many possible reasons why participants – intentionally or unintentionally – may or may not self-report their actual age.

A review of channel statistics for *vingtaine* and *cinquantaine* in 2011 indicated that an average of just over 80% of *vingtaine* participants and just over 90% of *cinquantaine* participants were connected to the chat server through an Internet Service Provider/IP address in France or Belgium. Although channel statistics do not provide information regarding date or place of birth, nationality, or other demographic features, it seems to be the case that most of these participants are probably native speakers of French.

4.3 Data coding

In order to take advantage of the types of procedures made available by the multivariate analysis software Goldvarb X, the following factor groups (i.e., independent variables) were selected for this study: Letter as Independent Variable 1; Diacritic as Independent Variable 2; Age Group as Independent Variable 3; and Month of Data Collection as Independent Variable 4. Since the main objective of this study was to understand in more detail the variable use of diacritics, each token (i.e., each instance) of diacritic nonuse was coded as Absent, and each token of diacritic use was coded as Present. This study does not take into account instances of a diacritic used when it should never be present. For example, although the first vowel in *effacer 'to erase'* is typically pronounced the same way that an *e* with an acute accent is pronounced, there is no acute accent on the *e* at the beginning of *effacer* (because it is followed by a geminate consonant). This is the same for words with a similar pattern (e.g., *efficace 'efficient'* and *essentiel 'essential'*). We did not include such tokens in the present study because they were so rare. We have observed over the past two decades that when such tokens do indeed appear in synchronous chat, they are often produced by a single participant who is repeating a specific (misspelled) word.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 3. Absence and presence of diacritics

Diacritic	Absent	Present
acute accent	388 (23.9%)	1,235 (76.1%)
cedilla	205 (40.8%)	297 (59.2%)
grave accent	591 (51.7%)	553 (48.3%)
circumflex accent	298 (68.0%)	140 (32.0%)
diaeresis	6 (75.0%)	2 (25.0%)
Total	**1,488 (40.1%)**	**2,227 (59.9%)**

For both of the replications in the present study, the authors identified relative tokens and coded them independently in order to compare the initial data to be used for the statistical analysis. This allowed us to resolve any discrepancies regarding the codes themselves or missing tokens by comparing the two sets of codes.

4.4 Results

A cursory review of the tables provided below indicates that some aspects of traditional communication in French are reproduced – to greater or lesser degrees – in electronic discourse. For example, in Table 3, it is clear that the acute accent and the grave accent are the most frequently used diacritics in French. This indicates that there is not some different paradigm that has developed in digital communication environments. However, there seems to be something peculiar about the use of the cedilla (see Table 4) since it is used more frequently than it is not used, yet its overall potential and actual use (i.e., the combined number of tokens) is relatively low. This means that there is not an across-the-board correlation linking the combined number of tokens and relatively high(er) use of a diacritic.

The results provided in Table 5 seem to suggest that there is very little difference between overall presence-absence rates for the *vingtaine* channel and those of the *cinquantaine* channel. However, even though there may be no noticeable difference when considering frequency tallies (see also Table 6), only a statistical analysis will be able to measure for a potentially significant difference between the two groups.

As explained above in the section on Data Coding, we used separate codes for Letter and Diacritic. This allowed us to look at the preliminary analysis from a different perspective, which had the potential to reshape our thinking about ways in which we could replicate (i.e., partially or completely) the study by Al-Rashdan (2010) or potentially other previous work on the variable use of diacritics. After reviewing several initial rounds of Varbrul results (including scatterplots that indicated an unacceptable amount of outliers), we returned to the use of Grapheme as Independent Variable 1, a combination of Letter and Diacritic (see Table 7).

After excluding the tokens with fewer than 20 occurrences in the corpus (i.e., *î*, *ï*, and *ö*), the GoldVarb X software produced an Input (i.e., corrected mean) of 61.2%

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 4. Absence and Presence of Diacritics: Letters

Letter	Diacritics Absent	Diacritics Present
e	735 (33.3%)	1,473 (66.7%)
c	205 (40.8%)	297 (59.2%)
a	442 (52.2%)	405 (47.8%)
u	44 (59.5%)	30 (40.5%)
i	14 (63.6%)	8 (36.4%)
o	48 (77.4%)	14 (22.6%)
Total	**1,488 (40.1%)**	**2,227 (59.9%)**

Table 5. Absence and Presence of Diacritics: Age Groups

Group	Absent	Present
Twenty-somethings	844 (41.4%)	1,196 (58.6%)
Fifty-somethings	644 (38.4%)	1,031 (61.6%)
Total	**1,488 (40.1%)**	**2,227 (59.9%)**

Table 6. Absence and Presence of Diacritics: Session of Participation

Period	Absent	Present
1 (May)	756 (43.3%)	988 (56.7%)
2 (June)	732 (37.1%)	1,239 (62.9%)
Total	**1,488 (40.1%)**	**2,227 (59.9%)**

presence of diacritics, which is quite close to the mean of 60.2%. (The mean of 59.9% presence of diacritics in Table 6 includes all coded tokens.) A scattergram of the data points is provided as Figure 1, in which the size of each square is proportional to the number of cells represented. This visual representation of the cells serves as one indicator of the fit of the data to the model; in this case, the fit appears to be good since most of the cells are located along or quite close to the regression line (Tagliamonte, 2006: 224).

There were too few tokens (i.e., fewer than 20 of each) of the graphemes *î* (*i*-circumflex), *ï* (*i*-diaresis), *ë* (*e*-diaeresis), and *û* (*u*-circumflex) in the corpus to include them in this VARBRUL analysis. Although an initial analysis including the 18 tokens of *î* (*i*-circumflex) and the 32 tokens of *û* (*u*-circumflex) produced factor weights that were almost identical (i.e., maximum difference was one one-hundredth of a percent) to those produced for the analysis presented in Table 8, it

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 7. Absence and Presence of Diacritics: Graphemes

Grapheme	Diacritic Absent	Diacritic Present	% of Corpus (*n*)	% Rank
é	388 (23.9%)	1,235 (76.1%)	43.7% (1,623)	1
ç	205 (40.8%)	297 (59.2%)	13.5% (502)	3
ï	2 (50.0%)	2 (50.0%)	0.1% (4)	11
à	412 (51.3%)	391 (48.7%)	21.6% (803)	2
è	156 (52.2%)	143 (47.8%)	8.0% (299)	4
ù	23 (54.8%)	19 (45.2%)	1.1% (42)	8
û	21 (65.6%)	11 (34.4%)	0.9% (32)	9
ê	187 (66.3%)	95 (33.7%)	7.6% (282)	5
î	12 (66.7%)	6 (33.3%)	0.5% (18)	10
â	30 (68.2%)	14 (31.8%)	1.2% (44)	7
ô	48 (77.4%)	14 (22.6%)	1.7% (62)	6
ë	4 (100%)	0 (0.0%)	0.1% (4)	12
Total	**1,488 (40.1%)**	**2,227 (59.9%)**	**100% (3,715)**	

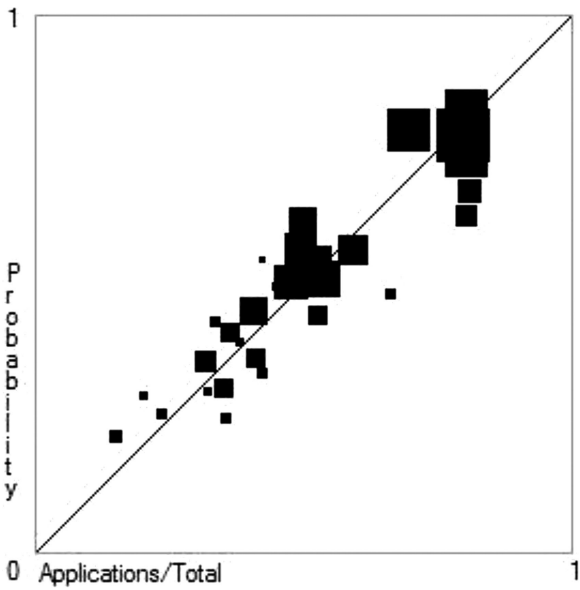


Figure 1. Scattergram of Cells from a Binomial One-Step Analysis (GoldVarb X).

became clear after examining the error values, log likelihood values, scattergrams, and other values that the analysis excluding *î* (*i*-circumflex) and *û* (*u*-circumflex) was a slightly better fit for this corpus (see Tagliamonte, 2006: 221-226; see also Bayley, 2002).

The results of the VARBRUL analysis of the recoded data are provided in Table 8. The main elements of interest are the weights and the ranges. Any weight above 0.50

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 8. VARBRUL Analysis of Recoded Data

Independent variable	Factor	No. present (and %)	GoldVarbX weight
Grapheme	é	1,235 (76.1%)	**.67**
	ç	297 (59.2%)	.49
	à	391 (48.7%)	.37
	è	143 (47.8%)	.36
	ù	19 (45.2%)	.35
	ê	95 (33.7%)	.24
	â	14 (31.8%)	.22
	ô	14 (22.6%)	.16
	û	———	———
	î	———	———
	ï	———	———
	ë	———	———
			Range: 51
Age	50-somethings	1,025 (62.4%)	**.53**
	20-somethings	1,183 (58.7%)	.47
			Range: 6
Period	1 (May)	981 (57.3%)	.47
	2 (June)	1,227 (63.1%)	**.52**
			Range: 5
Total no. present (and %)		**2,208 (60.4%)**	

(in bold) indicates that application of the rule (i.e., presence of a diacritic, in this case) is favoured, and a weight below 0.50 indicates that the application of the rule is disfavoured.[1] The results in Table 8 indicate that statistical significance was found for all three independent variables: grapheme, age, and period. Within these categories the weight is provided for each item (i.e., factor). For example, in the Grapheme category, *é* was found to be favoured, but the other graphemes were found to be disfavoured. In the Age category, the application of the rule was favoured in the data from the *cinquantaine* (50s) channel, and the application of the rule was also favoured in the June half of the corpus in the Period category. However, the ranges indicate that Grapheme exerted the greatest influence on the variable presence or absence of diacritics since this

[1]If no weights are provided in a table of results, this typically indicates that the statistical procedure did not find significance.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 9. Partitioned Data: Session 1 (May) vs. Session 2 (June)

1-May ($n = 981$)			2-June ($n = 1,227$)		
Diacritic	% Present	Gvarb X Weight	Diacritic	% Present	Gvarb X Weight
é	73.2	**.67**	é	78.2	**.67**
ç	57.7	**.50**	ç	60.9	.46
à	46.3	.39	à	51.0	.36
è	47.9	.40	è	47.7	.33
ù	34.8	.28	ù	57.9	.44
ê	35.9	.29	ê	31.8	.20
â	22.2	.17	â	38.5	.25
ô	16.2	.12	ô	32.0	.21
		Range: 54			Range: 46
Age Group	% Present	Gvarb X Weight	Age Group	% Present	Gvarb X Weight
20	56.0	[ ]	20	60.7	.46
50	58.6	[ ]	50	66.5	**.55**
		Range: [ ]			Range: 9

range was clearly the highest, and the other ranges were much lower, indicating that they influenced the variation to a much lesser extent.

The graphemes *î*, *ï*, *ë*, and *û* are not included in the data presented in Tables 9 and 10 since there were not enough tokens for partitioned analyses; cell counts at zero or close to zero can adversely affect the trustworthiness of the fit of the data to the statistical model (Tagliamonte, 2006: 209-212). Tables 9 and 10 serve only as an illustration of partitioning, which is another way to look at data in GoldVarb X. Data can be partitioned by running a VARBRUL analysis on specific parts of a corpus. For example, the May half of the corpus and the June half of the corpus in Study 1 were each run separately (see Table 9), then the *vingtaine* (20s) half of the corpus and the *cinquantaine* (50s) half of the corpus in Study 1 were each run separately (see Table 10). In Table 11, results from partitioned data including the three most frequently occurring graphemes are provided. In all three cases, we noticed that there was at least one (independent) variable that was not selected as significant by the statistical analysis. The purpose of looking at partitioned data is primarily to view the corpus, the data, and the results from different perspectives in case the partitioning shows what might appear to be an anomaly or a pattern that had gone unnoticed or unexplained.

4.5 Summary

Our primary motivation for undertaking Study 1 was to replicate Study 0 (Al-Rashdan, 2010). Even though these two studies were based on different data, both corpora are from the same larger corpus (i.e., collected during 2008 from the *vingtaine* (20s) and *cinquantaine* (50s) chat channels of EuropNet).

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 10. Partitioned Data: 20-somethings vs. 50-somethings

20-somethings			50-somethings		
Diacritic	% Present	Gvarb X Weight	Diacritic	% Present	Gvarb X Weight
é	78.2	**.70**	é	73.4	**.61**
ç	47.6	.38	ç	78.9	**.68**
à	42.9	.33	à	53.9	.41
è	47.4	.37	è	48.3	.35
ù	46.7	.37	ù	41.7	.30
ê	31.1	.23	ê	36.6	.25
â	30.8	.23	â	33.3	.22
ô	21.1	.15	ô	25.0	.16
Range: 55				Range: 52	
Session	% Present	Gvarb X Weight	Session	% Present	Gvarb X Weight
1-May	56.0	[ ]	1-May	58.6	.46
2-June	60.7	[ ]	2-June	66.5	**.54**
Range: —				Range: 8	

Table 11. VARBRUL analysis of the grapheme *é*, *à*, and *ç*

é		*à*		*Ç*	
Age group	Factor Weight	Age group	Factor Weight	Age group	Factor Weight
20	**.53**	20	.44	20	.37
50	.47	50	**.55**	50	**.71**
Range: 6		Range: 11		Range: 34	
Session	Factor Weight	Session	Factor Weight	Age group	Factor Weight
1-May	.47	1-May	[ ]	1-May	[ ]
2-June	**.53**	2-June	[ ]	2-June	[ ]
Range: 6		Range: ———		Range: ———	

There is a noticeable – yet not extreme – increase in the presence of diacritics. This is, of course, not surprising since it is very unlikely that two different chat sessions would produce exactly the same distribution. A closer comparison of Study 0 and Study 1 revealed some similarities and some additional differences (see Tables 12 and 13).

The comparison of the VARBRUL results for Study 0 and Study 1 shows first of all that Study 1 is not an exact replication of Study 0. However, this can be considered a replication since both factors from Study 0 were included in Study 1. Before we ran the

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 12. Distribution of tokens: presence and absence

Study	Tokens	Presence (number)	Presence (percentage)	Absence (number)	Absence (percentage)
0	3,855	2,136	55.4%	1,719	44.6%
1	3,715	2,227	59.9%	1,488	40.1%

Table 13. A Comparison of VARBRUL results: Study 0 and Study 1

Study	Significance	Range	Favored (presence)
0	Grapheme: Y	53	é, ç
	Age group: N	[ ]	[ ]
	Period: N/A	N/A	N/A
1	Grapheme: Y	51	é
	Age group: Y	6	*cinquantaine*
	Period: Y	5	June

final statistical procedure in GoldVarb X (Table 8), we included only the variables Grapheme and Age to be sure that when we added Period (May vs. June) as an experimental category, it would not skew the results. We ultimately decided to include Period as a variable since it did not interfere with the integrity of the analysis; however, this category probably would not be included in a traditional analysis of variation since we have no way of knowing if the May speech community or the June speech community was really two distinct communities or the same community or a community with a small or large amount of overlap. This issue ties into one of the main problems associated with analysing language variation in online communities like chat since biographical information (used for determining macrosociological features or variables) is not verifiable. This is certainly one of the main differences between traditional (i.e., written or spoken) data versus many types of digital data.

Another important aspect to point out in Table 13 is the difference between Study 0, which found no statistically significant difference between *vingtaine* and *cinquantaine*, and Study 1, which did indeed find a statistically significant difference between the two (self-reported) age groups. However, even though Study 1 found age to be statistically significant, it is clear that Grapheme influenced the application of the rule (i.e., the variable use of diacritics) much more than Age.

5. RESULTS FOR STUDY 2 – BW16 CORPUS

5.1 The corpus

The data for this study were taken from two real-time chat channels based on the age groups *vingtaine* 'twenty-something' and *cinquantaine* 'fifty-something' on

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 14. Word counts

Year	Vingtaine	Cinquantaine	Total
2008	29,310	31,054	**60,364**
2016	29,554	30,466	**60,020**
Total	**58,864**	**61,520**	**120,384**

EuropNet, a public federation of European Internet Relay Chat networks. These are the same chat channels that were used to create the 2008 corpus used for the study reported in the previous section. The corpus analysed below (i.e., half from 2008 and half from 2016) consists of several data sets that form two approximately equal halves of the corpus, whether it is divided by the year of data collection or the (perceived) age group of the participants (see Table 14).

5.2 Participants

As explained above in the section that provides a report of Study 1, little to nothing is known about synchronous chat participants. Even the scant amount of demographic information available is often unverifiable. However, when participants self-select a communication space based on an age group, this known characteristic can be isolated as an independent variable and used as an indication of belonging to a certain group (i.e., the *vingtaine* community or the *cinquantaine* community), even if these groups are not only composed of people in these two specific age ranges.

A review of channel statistics for *vingtaine* and *cinquantaine* indicated that an average of just over 80% of *vingtaine* participants and just over 90% of *cinquantaine* participants were connected to the chat server through an Internet Service Provider/IP address in France or Belgium. Although channel statistics do not provide information regarding date or place of birth, nationality, or other demographic features, it seems to be the case that most of these participants are probably native speakers of French.

5.3 Data coding

The following factor groups (i.e., independent variables) were selected for this study: Grapheme (defined for the purpose of this study as a letter with or without a diacritic) as Independent Variable 1; Age Group (i.e., perceived age group) as Independent Variable 2; and Year of Data Collection as Independent Variable 3. Since the main objective of this study is to understand in more detail the variable use of diacritics, each token (i.e., each instance) of diacritic non-use was coded as Absent, and each token of diacritic use was coded as Present. Presence or absence of a diacritic is the *yes*/*no* dependent variable for this study.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 15. Variable Use of Diacritics – 2008 and 2016 Combined

Grapheme	Total	Presence
é	3,237	2,410 (74.5%)
à	1,597	835 (52.3%)
ç	1,040	618 (59.4%)
è	629	334 (53.1%)
ê	598	236 (39.5%)
ô	141	42 (29.8%)
â	90	30 (33.3%)
û	84	22 (26.2%)
ù	80	33 (41.2%)
î	63	12 (19.0%)
ï	10	3 (30.0%)
ë	6	0 (0.0%)
Total	**7,575**	**4,575 (60.4%)**

5.4 Results

Table 15 is arranged according to the most to least frequently appearing tokens in the corpus, and the third column represents the number and percentage of the use (Total Presence) of each diacritic. All other tables are arranged in the same order (i.e., total number of tokens, regardless of presence/absence of a diacritic or GoldVarb X factor weight). Even though the Grapheme column shows each grapheme with a diacritic, the Total column represents a combination of the number of tokens with the diacritic and the number of tokens where the diacritic could have been used but was not. For example, the preposition *à* can also be typed as *a*, but *a* can also represent the third-person singular present indicative form of *avoir* 'to have'; however, when *a* was used as a verb, it was not included as a token of *a*-preposition (or *à*-preposition).

There are several aspects of the coding that must be reviewed during the various steps of a variable rule analysis. The most obvious problem that can prevent an analysis from being undertaken is the presence of a Singleton (i.e., one factor in a group) since each group must have at least two factors to compare (e.g., *Vingtaine* channel vs. *Cinquantaine* channel).

Another problem that will prevent an analysis from proceeding is the appearance of a Knock Out, which is any element within a factor group that exhibits no variation. In the present study, this was the case for the grapheme *ë* since the diacritic (dieresis) was absent in all six tokens. A Knock Out cannot be included in a variable rule analysis; therefore, *ë* was eliminated by performing a Recode in GoldVarb X before continuing. Once any Singletons and Knock Outs have been addressed, Tagliamonte (2006) recommends a close examination of cross-tabulation 'in order to spot overlap, interaction and non-orthogonality' (182). An example of cross-tabulation is provided in Table 16.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 16. Cross-Tabulation of Year and Age/Channel

		20 *n*	20 %	50 *n*	50 %	Total	Total %
2008	**Presence**	1,196	59	1,031	62	**2,227**	**60**
	Absence	841	41	643	38	**1,484**	**40**
	Total	2,037	100	1,674	100	**3,711**	**100**
2016	**Presence**	1,460	83	888	42	**2,348**	**61**
	Absence	298	17	1,212	58	**1,510**	**39**
	Total	1,758	100	2,100	100	**3,858**	**100**
Total	**Presence**	2,656	70	1,919	51	**4,575**	**60**
	Absence	1,139	30	1,855	49	**2,994**	**40**
	Total	3,795	100	3,774	100	**7,569**	**100**

Note that *20* refers to the *Vingtaine* channel, and *50* represents the *Cinquantaine* channel.

The next step requires running a one-level binomial analysis, at which point scattergrams are produced as a visual representation of the cells along a regression line. For the present study, this analysis was performed without *ë*, since this grapheme has already been excluded. Figure 2 shows two clear outliers, which have been circled. These outliers represent cells with the grapheme *ï*. The outlier on the left represents a cell with three tokens of [*ï Vingtaine* 2016], and the outlier on the right represents a cell with one token of [*ï Cinquantaine* 2008]. Since these cells are clear outliers, and since the grapheme *ï* represents only 0.1% of the corpus, this grapheme has been excluded from the present study.

All apparent problems related to unwanted interaction have been resolved by excluding *ë* and *ï* (Recode 2 in GoldVarb X) at this stage of the variable rule analysis (see Tagliamonte, 2006: 180-190). The scattergram for Recode 2 showed that there were no longer any clear outliers, and the Log Likelihood moved closer to zero in Recode 2, which 'indicates that this run is a better fit of the model to the data' (Tagliamonte: 223).

In order to test for statistical significance, a step-up/step-down analysis was run using Recode 2, which excluded *ë* and *ï*. This type of regression analysis 'has become the "gold standard" in variationist sociolinguistics' (Tagliamonte, 2006: 226). Although measures have already been taken to remove unwanted interaction from the analysis, one final check was made by examining the results of the step-up/step-down analysis before accepting Recode 2 as acceptable cells for this procedure (see Tagliamonte: 229-234). The results of the step-up/step-down analysis of 7,559 tokens of the variable use of diacritics are provided in Table 17. The overall presence of diacritics (i.e., application of the rule in Varbrul terms) is 4,572 instances (60.5%) out of 7,559; the overall absence of diacritics (i.e., non-application of the rule in Varbrul terms) is 2,987 instances out of 7,559.

Although a variable rule analysis of all data combined can be useful to see the overall trends, partitioning the data by age group (*vingtaine* only and *cinquantaine* only) and year (2008 only and 2016 only) provides a more detailed representation of the variable use of diacritics.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

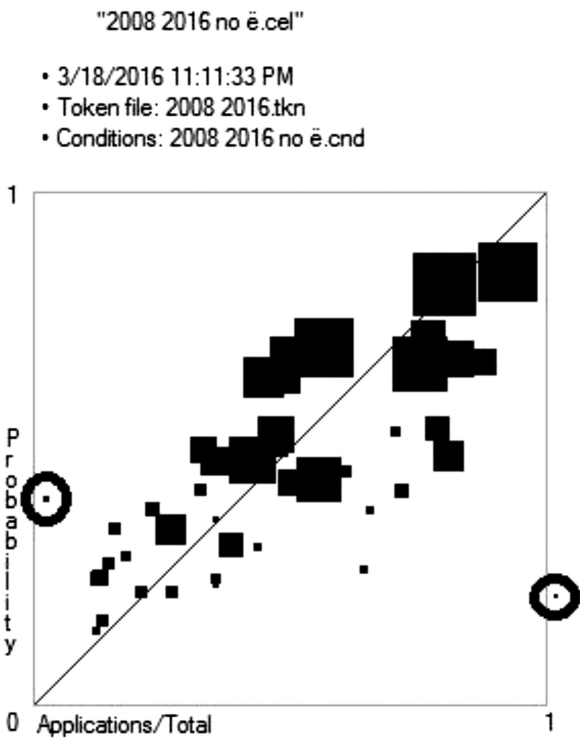


Figure 2. Scattergram from Varbrul Analysis: Recode 1.

Overall, the variable use (presence) of the graphemes *é*, *à*, and *ç*[2] remained quite stable between 2008 and 2016, as shown in Table 18 (see also Table 19 for the two age groups). However, it is important to note that this table does not demonstrate how each age group/channel converged or diverged regarding the variable use of diacritics. Tables 20, 21, 22, and 23 provide partitioned data that show results within a single factor (2008, 2016, *vingtaine* channel, *cinquantaine* channel, respectively).

5.5 Summary

The most fascinating aspect of these partition analyses comparing data from 2008 and data from 2016 is that the application of the rule (i.e., the presence of diacritics) has changed over time, at least in these two chat channels on the EuropNet server. In 2008, it is clear that the use of diacritics was favoured by *cinquantaine* 'people-in-their-fifties' participants, but by 2016, the use of diacritics was favoured by *vingtaine* 'people-in-their-twenties' participants. Trying to determine why this happened (and whether or not this might be a generalizable finding) would be purely speculative.

[2]All tokens (i.e., with or without diacritics) of *é*, *à*, and *ç* represent 77.5% of all the tokens analysed for Study 2.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 17. Variable Rule Analysis of Diacritics

Independent Variable	Factor	Presence *n* (%)	GoldVarbX Weight
Grapheme (Range: 52)	É	2,410 (74.5%)	**.65**
	À	835 (52.3%)	.42
	Ç	618 (59.4%)	.47
	È	334 (53.1%)	.42
	Ê	236 (39.5%)	.29
	Ù	33 (41.2%)	.29
	Â	30 (33.3%)	.23
	Ô	42 (29.8%)	.20
	Û	22 (26.2%)	.20
	Î	12 (19.0%)	.13
	Ï	[excluded]	
	Ë	[excluded]	
Age (Range: 20)	*Vingtaine*	2,655 (70.1%)	**.60**
	Cinquantaine	1,917 (50.8%)	.40
Year (Range: 4)	2008	2,225 (60.0%)	.48
	2016	2,347 (60.9%)	**.52**

Table 18. Variable Use of Diacritics: 2008 and 2016

Grapheme	2008 Total	2008 Presence	2016 Total	2016 Presence
é	1,623	1,235 (75%)	1,614	1,175 (73%)
à	803	391 (49%)	794	444 (56%)
ç	502	297 (59%)	538	321 (60%)
è	299	143 (48%)	330	191 (58%)
ê	282	95 (34%)	316	141 (45%)
ô	62	14 (23%)	79	28 (35%)
â	44	14 (32%)	46	16 (35%)
û	32	11 (34%)	52	11 (21%)
ù	42	19 (45%)	38	14 (37%)
î	18	6 (33%)	45	6 (13%)
ï	4	2 (50%)	6	1 (17%)
ë	4	0 (0%)	2	0 (0%)
Total	**3,715**	**2,227 (60%)**	**3,860**	**2,348 (61%)**

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 19. Variable Use of Diacritics: Vingtaine (20) and Cinquantaine (50)

Grapheme	20 Total	20 Presence	50 Total	50 Presence
é	1,733	1,457 (84.1%)	1,504	953 (63.4%)
à	685	409 (59.7%)	912	426 (56.7%)
ç	577	346 (60.0%)	463	272 (58.7%)
è	306	204 (66.7%)	323	130 (40.2%)
ê	272	140 (51.5%)	326	96 (29.4%)
ô	71	31 (43.7%)	70	11 (15.7%)
â	48	21 (43.8%)	42	9 (21.4%)
û	24	14 (58.3%)	60	8 (13.3%)
ù	46	25 (54.3%)	34	8 (23.5%)
î	27	8 (29.6%)	36	4 (11.1%)
ï	6	1 (16.7%)	4	2 (50.0%)
ë	3	0 (0.0%)	3	0 (0.0%)
Total	**2,040**	**1,196 (59%)**	**1,758**	**1,460 (83%)**

Table 20. Partition Analysis: 2008

Independent Variable	Factor	Presence *n* (%)	GoldVarb X Weight
Grapheme (Range: 51)	É	1,235 (76.1%)	**.67**
	À	391 (48.7%)	.37
	Ç	297 (59.2%)	.48
	È	143 (47.8%)	.37
	Ê	95 (33.7%)	.24
	Ù	19 (45.2%)	.35
	Â	14 (31.8%)	.23
	Ô	14 (22.6%)	.16
	Û	11 (34.4%)	.24
	Î	[excluded]	
	Ï	[excluded]	
	Ë	[excluded]	
Channel (Range: 5)	*Vingtaine*	1,190 (58.7%)	.48
	Cinquantaine	1,029 (61.9%)	**.53**

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 21. Partition Analysis: 2016

Independent Variable	Factor	Presence *n* (%)	GoldVarb X Weight
Grapheme (Range: 47)	É	1,175 (72.8%)	**.62**
	À	444 (55.9%)	.46
	Ç	321 (59.7%)	.46
	È	191 (57.95%)	.44
	Ê	191 (57.9%)	.33
	Ù	14 (36.8%)	.23
	Â	16 (34.8%)	.19
	Ô	28 (35.4%)	.22
	Û	11 (21.2%)	.15
	Î	[excluded]	
	Ï	[excluded]	
	Ë	[excluded]	
Channel (Range: 46)	*Vingtaine*	1,457 (83.9%)	**.75**
	Cinquantaine	884 (42.7%)	.29

Table 22. Partition Analysis: Vingtaine Channel

Independent Variable	Factor	Presence *n* (%)	GoldVarb X Weight
Grapheme (Range: 47)	É	1,457 (84.1%)	**.68**
	À	409 (59.7%)	.36
	Ç	346 (60.0%)	.36
	È	204 (66.7%)	.42
	Ê	140 (51.5%)	.28
	Ù	25 (54.3%)	.33
	Â	21 (43.8%)	.21
	Ô	31 (43.7%)	.21
	Û	14 (58.3%)	.31
	Î	[excluded]	
	Ï	[excluded]	
	Ë	[excluded]	
Year (Range: 34)	2008	1,190 (58.7%)	.34
	2016	1,457 (83.9%)	**.68**

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Table 23. Partition Analysis: Cinquantaine Channel

Independent Variable	Factor	Presence *n* (%)	GoldVarb X Weight
Grapheme (Range: 50)	É	953 (63.4%)	**.63**
	Ç	426 (46.7%)	.45
	À	272 (58.7%)	.59
	È	130 (40.2%)	.39
	Ê	96 (29.4%)	.29
	Ù	8 (23.5%)	.24
	Â	9 (21.4%)	.20
	Ô	11 (15.7%)	.16
	Û	8 (13.3%)	.13
	Î	[excluded]	
	Ï	[excluded]	
	Ë	[excluded]	
Year (Range: 20)	2008	1,029 (61.9%)	**.61**
	2016	884 (42.7%)	.41

6. CONCLUSION

In this study, we have highlighted the potential value of replicating research that explores sociolinguistic variation by demonstrating that expanding a corpus with an additional data set can substantially alter the findings, and comparisons over time can provide an overview of shifting communicative practices. Another type of replication could include data from the same corpus that is analyzed using different software or including additional factors. In other words, replication is not limited to a diachronic analysis.

Our replications demonstrated – among other things – that data sets taken from different parts of the same corpus do not necessarily show the same or similar results. In Al-Rashdan's (2010) study, age was not found to be a factor that influenced the variable use of diacritics, but in Study 1 (our replication of Al-Rashdan's study), the use of diacritics was shown to be statistically favoured by the *cinquantaine* (50s) group. However, in Study 2 (our replication of Study 1), our diachronic analysis revealed that almost a decade later, the use of diacritics was shown to be statistically favoured by the *vingtaine* (20s) group. Although we feel strongly that the development of new devices (e.g., smartphones, smartwatches, tablets) and the improvement of different technologies (e.g., auto-correction, voice recognition software) might influence trends related to the variable use of diacritics, this is only speculation on our part.

Our second replication of Al-Rashdan's study has indeed demonstrated a clear change over time; however, the sole reliance on evidence from a statistical analysis

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

led us to begin thinking more seriously about the potential value of adding a qualitative dimension. When we began to imagine how excerpts could be used to add a qualitative dimension to any analysis of computer-mediated discourse, we were reminded of the lack of verifiable biographic data of the participants, which is a limitation encountered by most researchers whose work is based on data from online sources. In spite of certain limitations, it is clear that exploring online discourse is certainly a worthwhile endeavour.

Funding. There are no funders to report for this submission

References

Académie française. (n.d.). Rectifications de l'orthographe. URL: http://www.academie-francaise.fr, retrieved 12 June 2020.

Al-Rashdan, O. (2010). L'emploi variables des signes diacritiques dans le français tchaté : une étude variationniste en temps apparent. Master's thesis, University of North Texas.

Anis, J. (1988). *L'écriture: théories et descriptions*. Bruxelles: De Boeck.

Anis, J. (1999). Chat et usages graphiques. In : J. Anis (ed), *Internet, communication et langue francaise*. Paris: Hermès, pp. 71–90.

Bayley, R. (2004). The quantitative paradigm. In: J. K. Chambers, P. Trudgill and N. Schilling-Estes (eds), *The Handbook of Language Variation and Change*. Malden, MA: Blackwell, pp. 127–128.

Bauman, R. (1992). Performance. In: R. Bauman (ed), *Folklore, Cultural Performances, and Popular Entertainments*. New York: Oxford University Press, pp. 41–49.

Bourdieu, P. (1977). The economics of linguistic exchanges. *Social Science Information*, **16**: 645–668.

Catach, N. (1973). La structure de l'orthographe française. *La Recherche*, **39**: 949–956.

Catach, N. (1980). *L'orthographe française*. Paris: Nathan.

Catach, N. (1991). *Orthographe ! Orthographe !* Paris: Duculot.

Catach, N., Gruaz, C. and Duprez, D. (1986). *L'orthographe française: traité théorique et pratique*. Paris: Nathan.

Cougnon, L.-A. (2010). Orthographe et langue dans les SMS: conclusions à partir de quatre corpus francophones. *Études de Linguistique Appliquée*, **160**: 397–410.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Danet, B., Ruedenberg-Wright, L. and Rosenbaum-Tamari, Y. (1997). 'Hmmm…Where's that smoke coming from?': Writing, play and performance on Internet Relay Chat. Journal of Computer-Mediated Communication, 2.4: n.p. URL: https://academic.oup.com/jcmc/article/2/4/JCMC246/4584364, retrieved 16 July 2021.

Daussy, L. (2009). La réforme de l'orthographe mieux appliquée à l'étranger. Le Figaro. URL: https://www.lefigaro.fr/actualite-france/2009/12/21/01016-20091221ARTFIG00660-la-reforme-de-l-orthographe-mieux-apliquee-a-l-etranger-.php, retrieved 11 July 2021.

EuropNet Statistics. URL: http://stats.europnet.org/, retrieved 2 February 2018.

Fayol, M. and Jaffré, J.-P. (2016). L'orthographe: des systèmes aux usages. *Pratiques*, **169**: 1–15.

Finnegan, R. (1992). *Oral Traditions and the Verbal Arts*. London: Routledge.

Gadet, F. (2008). Ubi scripta et volant et manent. In: E. Stark, R. Schmidt-Riese and E. Stoll (eds), *Romanische Syntax im Wandel*. Tübingen: Narr, pp. 513–529.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. Language@Internet, 4, 1–37. URL: http://www.languageatinternet.org/articles/2007/761, retrieved 2 January 2015.

Herring, S. C. and Zelenkauskaite, A. (2009). Symbolic capital in a virtual heterosexual market: Abbreviation and insertion in Italian iTV SMS. *Written Communication*, **26**: 5–31.

Labov, W. (1963). The social motivation of a sound change. *Word*, **19**: 273–309.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Myslín, M. and Gries, S. Th. (2010). k dixez? A corpus study of Spanish Internet orthography. *Literary and Linguistic Computing*, **25**: 85–105.

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press

Pellat, J-C. and Andrieux-Reix, N. (2006). Histoire d'E ou de la variation des usages graphiques à la différentiation réglée. *Langue Française*, **151**: 7–24.

Pierozak, I. (2000). Approche sociolinguistique des pratiques discursives sur internet: 'ge fé dais fotes si je vœux'. *Revue française de linguistique appliquée*, **5** : 89–104.

Pierozak, I. (2003). Le francais tchaté :Une étude en trois dimensions – sociolinguistique, syntaxique et graphique – d'usage IRC. Doctoral dissertation, Université d'Aix-Maseille I, France.

Rickard, P. (1993). *A History of the French Language*, 2nd edn. London: Routledge.

Stark, E. (2015). 'De l'oral dans l'écrit'? – Le profil variationnel des SMS (textos) et leur valeur pour la recherche linguistique. In: K. Jeppesen Kragh and J. Lindschouw (eds), *Les variations diasystématiques et leurs interdépendances dans les langues romanes, Actes du Colloque DIA II à Copenhague*. Strasbourg: ELiPhi, pp. 395–405.

Sykes, J. (2005). Synchronous CMC and pragmatic development: Effects of oral and written chat. *CALICO Journal*, **22**: 399–421.

Tagliamonte, S. (2006). *Analysing Sociolinguistic Variation*. Cambridge, UK: Cambridge University Press.

Tagliamonte, S. (2012). *Variationist Sociolinguistics: Change, Observation, Interpretation*. Malden, MA: Wiley-Blackwell.

Tatossian, A. (2008). Typologie des procédés scripturaux des salons de clavardage en français chez les adolescents et les adultes. In: J. Durand, B. Habert and B. Laks (eds), *Actes du 1er Congrès mondial de linguistique française*. Paris: Institut de Linguistique Française, pp. 2337–2352.

van Compernolle, R. A. (2011). Use and variation of French diacritics on an Internet dating site. *Journal of French Language Studies*, **21**: 131–148.

van Compernolle, R. A. and Williams, L. (2007). De l'oral à l'électronique: La variation orthographique comme ressource sociostylistique et pragmatique dans le français électronique. *Glottopol*, **10**: 56–69.

van Compernolle, R. A. and Williams, L. (2010). Orthographic variation in electronic French: The case of l'accent aigu. *French Review*, **83**: 820–833.

Watts, F. (1991). Reflexions sur la proposition de reforme de l'orthographe et sa polemique. *The French Review*, **65**: 84–90.

Werry, C. (1996). Linguistic and interactional features of Internet relay chat. In: S. C. Herring (ed), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural perspectives*. Phildelphia: Benjamins, pp. 47–63.

Williams, L. (2009). Sociolinguistic variation in French computer-mediated communication: A variable rule analysis of the negative particle ne. *International Journal of Corpus Linguistics*, **14**: 67–491.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. In: S. C. Herring (ed), *Computer-mediated communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam: John Benjamins, pp. 29–46.

Yus, F. (2010). *Ciberpragmática 2.0. Nuevos Usos del Lenguaje en Internet*. Barcelona: Ariel.

Zelenkauskaite, A. (2017). Abbreviate and insert? Message length„addressee and non-standard writing in Italian mobile texting and Facebook. *Writing Systems Research*, **9**: 123–136.

Zelenkauskaite, A. and Gonzales, A. L. (2017). Non-standard typography use over time: Signs of a lack of literacy or symbolic capital? *The Journal of Community Informatics*, **13**: 72–91.

Zelenkauskaite, A. and Herring, S. C. (2006). Gender encoding of typographical elements in Lithuanian and Croatian IRC. In: F. Sudweeks, H. Hrachovec and C. Ess (eds), *Cultural Attitudes towards Technology and Communication 2006*. Murdoch, Australia: School of Information Technology, Murdoch University, pp. 474–489.

Cite this article: Blattner G and Williams L (2022). The variable use of diacritics in synchronous computer-mediated French discourse: Replication research. *Journal of French Language Studies* **32**, 216–242. https://doi.org/10.1017/S0959269522000011

https://doi.org/10.1017/S0959269522000011 Published online by Cambridge University Press