# The Dutch verb-spelling paradox in social media

## A corpus study

Tijn Schmitz, Robert Chamalaun and Mirjam Ernestus
Radboud University

Although the Dutch verb spelling system seems very straightforward, many spelling errors are made, both by children and adults (e.g., Sandra, Frisson, & Daems 2004). These errors mainly occur with verbs with two or more homophonous forms in their inflectional paradigms. Ample experimental research has been carried out on this topic, but these studies hardly reflect everyday language behavior. In the current corpus study, we reassessed previously found experimental results, but now in a Twitter corpus containing 17,432 tweets with homophonous verb forms. In accordance with previous results, we found a clear preference for the suffix -<*d*> compared to both -<*dt*> and -<*t*>, as well as a frequency effect, resulting in fewer errors for more frequent word forms. Furthermore, the results revealed that users with more followers make fewer errors, and that more errors are made during the evening and night.

**Keywords:** verb spelling errors, homophones, frequency effect, mental lexicon, everyday language behavior

### 1. Introduction

For theories of language production, spelling errors are an informative phenomenon. Studying spelling errors can be helpful in making inferences regarding principles underlying lexical representation and morphological and phonological processing. Homophones represent an especially interesting situation. Previous experimental work on Dutch and French showed that when a single pronunciation is spelled differently depending on its grammatical function, the spelling process of the correct form can be impeded (e.g., Sandra & Fayol 2003). The current study investigates which factors play a role in spelling errors in homophones in everyday language behavior.

*Linguistics in the Netherlands 2018*, **111–124**. https://doi.org/10.1075/avt.00008.sch
ISSN 0929–7332 / E-ISSN 1569–9919 

In languages such as Dutch and French, homophonous inflected verb forms lead to an intriguing paradox. On the one hand, the descriptive complexity of the spelling rules is low, as they can largely be characterized as morphographic (concatenating the stem and one or multiple suffixes). As a consequence, the spelling of inflected verb forms does not only reflect their pronunciation but also their morphological structure. On the other hand, many errors are made, even by experienced writers. Sandra (2010) showed that Dutch 18-year-olds make up to 25% errors in spelling homophonous verb forms. Given the relative simplicity of the rules, the persistence of these errors suggests that they are caused by the nature of cognitive processes underlying the spelling process, rather than negligence or laziness.

Previous studies suggest that writers do not always determine the spelling of a verb form by applying the spelling rules (the computational procedure; Sandra, Frisson, & Daems 1999; Verhaert, Danckaert, & Sandra 2016). Writers often retrieve the spellings of verb forms from their mental lexicons (Sandra & van Abbenyen 2009). This retrieval procedure works well for non-homophonous words, but presents problems for homophonous words when it is only based on sound-spelling correspondence, not taking the grammatical function of the verb form into account. According to several models (e.g., Baayen, Dijkstra, & Schreuder 1997; Laudanna & Burani 1985), the computational procedure and the retrieval procedure are in competition and the procedure that is fastest determines the outcome.

The speed of the two procedures is determined by several important factors. The first factor is the frequency of occurrence of the base of the verb form, for the computational procedure. The second factor is the frequency of occurrence of the spelling of the verb form itself, for the retrieval procedure. The higher these frequencies are, the faster the procedures can be applied. For the retrieval of homophonous forms, the *relative* frequency of the spelling of the target homophone relative to the other homophone is also relevant. Retrieval of the most frequent spelling is easier and the retrieval procedure therefore prefers the most frequent spelling, even if that form is ungrammatical according to the intended grammatical function.

Another important factor determining the ease of the spelling process is the distance between the verb form and the word determining its suffix. For present tense verbs, the suffix is determined by the sentence's grammatical subject, while for participles the auxiliary verb is important. When the distance to the word determining the verb's suffix increases, so does the probability of an error (e.g., Fayol, Largy, & Lemaire 1994; Largy & Fayol 1996; Sandra et al., 2004). When verb form and the word determining its suffix are adjacent (as in *hij betaalt* "he pays" and *hij heeft betaald* "he has paid"), the information needed to select the verb

suffix is still salient in working memory. By contrast, when verb form and the word determining its suffix are separated by one or multiple words (as in *dat hij mij morgen betaalt* "that he will pay me tomorrow" and *hij heeft mij nog niet betaald* "he has not yet paid me"), the spelling of these other words may interfere with the information stored in working memory for determining the verb form, making it harder to determine the correct output via the computational procedure.

Most studies on Dutch verb spelling involve experiments with error-evoking contexts (e.g., Bosman 2005; Sandra et al., 1999; Verhaert 2016), in which the participant's task is, for instance, to insert a verb form in a sentence or metalinguistic tasks such as indicating which strategy they used to determine the verb form. Additionally, time pressure leads to an increased working load, which has been shown to increase the number of errors (Fayol et al., 1994). Most participants in previous experiments were university or high school students, which is a rather limited reflection of the entire population. Furthermore, as participants are often aware that the study is about spelling, they might use different strategies than they would do in spontaneous writing. These experimental studies do not reflect spontaneous production of complete sentences and the question arises whether their results are ecologically valid.

A corpus study seems an obvious first step in this line of research, but so far, no real corpus studies have been conducted on this topic (however, see Sandra 2010). The current study sheds new light on Dutch verb spelling errors using a Twitter corpus. In contrast to artificial testing situations, tweets are spontaneous utterances. Usually, twitterers are busy with their everyday life while writing a tweet and do not think too long about what they write, which contributes to the spontaneous character of tweets. Furthermore, twitterers come from all ranges of the population: young, old, lower and higher educated (van der Veer, Boekee, & Peters 2017), making it more justified to generalize results to a larger population. Thus, Twitter is an ideal medium to investigate how spelling rules are maintained in spontaneous language production of a larger population.

The most important factors found to induce spelling errors in previous studies – relative frequency and adjacency – will again be assessed in this corpus study to investigate whether they also affect spontaneous language production. Furthermore, this study provides an excellent opportunity to investigate several factors that are specific to Twitter, such as the time of day the tweet was sent, the tweet length, and the number of followers. A study by Hutto, Yardi, and Gilbert (2013) showed that Twitter users seek out well-written content over poorly written content when deciding whether to follow another user. As people with better general language skills tend to be better spellers, we expect that the higher the number of followers, the fewer mistakes in spelling.

We focus on two Dutch verb spelling rules, associated with two types of homophones. The first rule, illustrated in Example (1), marks second and third person singular (henceforth referred to as "third person singular"), and adds -<*t*> to the stem.

(1) *Hij werk+t* /ɦɛi ʋɛrk+t/ 'He works'

In verbs with stem-final -<d>, application of this rule leads to a homophone pair: As illustrated in (2) for the verbal stem *word*, these verbs have a form ending in -<*d*> for the first person singular and -<*dt*> for the third person singular. Due to final devoicing and degemination, both verb forms are pronounced identically, namely as ending in /t/.

(2) a. *Ik word* /ɪk ʋor**t**/ 'I become'
b. *Hij word+t* /ɦɛi ʋor**t**/ 'He becomes'

The second rule marks the past participle and adds the prefix <*ge*>- and either the suffix -<*d*> or -<*t*>, depending on the last consonant of the stem. When this sound is voiceless, as the /*k*/ in *werk* in (3a), the suffix is -<*t*>; when it is voiced, as the /*m*/ in *noem* in (3b), the suffix is -<*d*>.

(3) a. *ge+werk+t* /xə+ʋɛrk+t/ 'worked'
b. *ge+noem+d* /xə+num+t/ 'mentioned'

The homophone type associated with this rule occurs in so-called *weak-prefix-verbs*. These are verbs starting with an unstressed (semi-)prefix (e.g., /**x**əbør/ 'happen', /**b**əlof/ 'promise'). The past participles of these verbs do not have an additional prefix *ge*-, leading to a homophone pair of the third person singular and past participle, as illustrated in (4). Due to final devoicing, both verb forms are pronounced as ending in /t/.

(4) a. *Het gebeur+t* /ɦɛt xəbør+**t**/ 'It happens'
b. *Het is gebeur+d* /ɦɛt ɪs xəbør+**t**/ 'It has happened'

## 2. Method

### 2.1 The corpus

Our corpus was extracted from TwiNL, a database of Dutch tweets posted from December 2010 onwards (Tjong Kim Sang & van den Bosch 2013). We used all 711,022 tweets from September 4, 2017. The character limit was then still 140 characters. All retweets were excluded, leaving 470,280 unique tweets.

Manual annotation of the correctness of verb forms in 2,500 tweets revealed that not all tweets were usable. For instance, tweets that were clearly produced by Twitterbots, automatically translated tweets, and standard patterns generated by apps (e.g., *Ik vind een Youtube-video leuk* 'I like a Youtube video') are not spontaneously produced but rather pre-programmed standard messages. This makes them unusable for the current study, as we focus on spontaneous language production. With a computer program, we excluded these unusable tweets. The estimated accuracy of the exclusion procedure was minimally 99%, based on a sample of 100 excluded tweets, meaning that (almost) all excluded tweets were genuinely unusable. After this procedure, 339,829 tweets remained. A sample of 100 of these remaining tweets showed that estimated accuracy of inclusion was minimally 99% as well, meaning that (almost) all tweets that were not excluded were genuinely usable.

We selected all tweets containing at least one homophonous verb form ending in -<*d*> , -<*dt*> , or -<*t*>, using the Dutch Morphology Wordforms list from CELEX (Baayen, Piepenbrock, & Van Rijn 1995). When the tweet contained multiple homophonous verb forms, these were assessed separately. As the tweets were not part-of-speech-tagged, it was possible that the selected homophone functioned as a noun (e.g., *antwoord* 'answer') instead of a verb form. When a form more frequently functioned as a noun than as a verb form, the homophone pair was discarded. Furthermore, tweets lacking a subject were discarded as well, leading to a final corpus of 17,432 tweets.

We then developed a program to classify all verb forms as correct/incorrect. As the tweets were not part-of-speech-tagged, the subject of the phrase was not known, nor was it known whether the verb form was a main or auxiliary verb. The program thus incorporated multiple search patterns, as the verb forms could be used in various constructions, which all represented unique patterns. In this process, both the type-I-error (classifying a correct verb form as incorrect) and type-II-error (classifying an incorrect verb form as correct) had to remain as low as possible. As a criterion, we maintained a threshold value of 90% accuracy for each search pattern, which was calculated based on a sample of 50 hits per pattern. The final program consisted of 38 search patterns (21 for errors, 17 for correctly-spelled verb forms). The search pattern accuracy varied from 90% to 100%, with a weighted average of 97.2% for all search patterns. A remainder of 276 verb forms could not be classified by the program and was annotated manually.

### **2.2** Analysis

We analyzed the data with Logistic Regression (stepwise, both directions). We conducted separate analyses for Homophone Type 1 (first vs. third person

singular) and Homophone Type 2 (third person singular vs. past participle). The verbs *worden* and *vinden* were excluded from the main analyses because of their high frequencies, and were analyzed separately. The dependent variable for all analyses was *score*, the probability that the spelling of a verb form was incorrect. The independent variables were:

– *Relative frequency* (log-transformed) of the correct form compared to its homophone counterpart. Frequencies were taken from the Dutch Morphology Wordforms from CELEX. Relative frequency was calculated using the following formula:

$$relative\ frequency = \log\left(\frac{frequency_{correct} + 1}{frequency_{homophone} + 1} + 1\right)$$

– *Correct suffix*. In the analyses of Homophone Type 1, *worden*, and *vinden*, the correct suffix was either -<*d*> or -<*dt*> . In the analysis of Homophone Type 2, it was either -<*d*> or -<*t*> .
– *Adjacency*. Binary variable (yes/no) indicating whether or not the verb form and the word determining its suffix were adjacent.
– *Time*. Tweets were categorized in four time groups: morning (6:00h-12:00h), afternoon (12:00h-18:00h), evening (18:00h-0:00h), and night (0:00h-6:00h).
– *Number of followers* (log-transformed).
– *Tweet length* (number of characters).

## 3. Results

Ultimately, 6.8% of all verb forms in the corpus were found to be incorrectly spelled. We will discuss this result in the Discussion section.

### 3.1 Homophone pair 1 (first vs. third person singular)

The results of the statistical analysis of Homophone Pair 1 are summarized in Table 1.

First, several effects were found of the variables investigated in previous experimental studies. When the correct suffix was -<*dt*> , significantly more errors were made than when the correct suffix was -<*d*> . The simple effect of *Relative Frequency* and its interaction with *Adjacency* showed that a higher relative frequency led to fewer errors. Unexpectedly, this was especially so when verb form and the word determining its suffix were not adjacent.

**Table 1.** Coefficients of Logistic regression on Homophone Pair 1 (*-d* vs. *-dt*), predicting incorrectness of target verb form, with higher *β*-values corresponding to a higher error rate

| Variable | *β* | Z-value | P-value |
|---|---|---|---|
| (Intercept) | −3.788 | −7.110 | <.001 |
| Relative Frequency | −0.284 | −1.930 | <.01 |
| Correct suffix | 3.181 | 8.865 | <.001 |
| Adjacency | −0.140 | −0.417 | >.1 |
| Time: Afternoon | 0.086 | 0.439 | >.1 |
| Time: Evening | 0.619 | 3.106 | <.01 |
| Time: Night | 0.826 | 2.435 | <.05 |
| Followers | −0.208 | −5.606 | <.001 |
| Tweet length | 0.006 | 2.336 | <.05 |
| Relative Frequency*Adjacency | −0.301 | −1.982 | <.05 |

Secondly, several effects were found of the Twitter-specific factors. An effect of *Time* showed that, with *Morning* as baseline, significantly more errors were made in tweets sent during the evening and night. No difference in number of errors was found when afternoon was compared to morning, nor did we find differences among afternoon, evening, and night. Furthermore, the analysis revealed an effect of *Followers*: Tweets written by users with more followers contained fewer errors. *Tweet length* turned out to have a significant effect as well: The longer the tweet, the more errors were made.

### 3.2 Homophone pair 2 (third person singular vs. past participle)

The results for the analysis of Homophone Pair 2 were largely comparable to those of Homophone Pair 1, and are summarized in Table 2.

As in previous studies, when the correct suffix was -<*t*>, significantly more errors were made than when the correct suffix was -<*d*>. Furthermore, *Relative Frequency* and *Adjacency* showed the same simple and interaction effects as for Homophone Pair 1.

The effect of *Time* was quantitatively slightly different from the analysis of Homophone Pair 1. With *Morning* as baseline, significantly more errors were made in tweets sent during the afternoon, evening, and night. No differences were found among afternoon, evening, and night. The analysis again revealed an effect of *Followers*: Users with more followers made fewer errors. Finally, *Tweet length* appeared not to be significant in this analysis.

**Table 2.** Coefficients of Logistic regression on Homophone Pair 2 (*-d* vs. *-t*), predicting incorrectness of target verb form, with higher *β*-values corresponding to a higher error rate

| Variable | *β* | Z-value | P-value |
|---|---|---|---|
| (Intercept) | −2.361 | −8.861 | <.001 |
| Relative Frequency | −0.283 | −3.265 | <.01 |
| Correct suffix | 1.090 | 6.152 | <.001 |
| Adjacency | −0.745 | −5.198 | <.001 |
| Time: Afternoon | 0.410 | 2.966 | <.01 |
| Time: Evening | 0.549 | 3.725 | <.001 |
| Time: Night | 0.635 | 2.475 | <.05 |
| Followers | −0.124 | −4.974 | <.001 |
| Tweet length | −0.001 | −0.340 | >.1 |
| Relative Frequency*Adjacency | −0.317 | −2.300 | <.05 |

### 3.3 *Worden*

The results of the analysis for *worden* are summarized in Table 3. *Relative Frequency* was omitted from this analysis, as the analysis contained only one verb.

**Table 3.** Coefficients of Logistic regression on *Worden* (*-d* vs. *-dt*), predicting incorrectness of target verb form, with higher *β*-values corresponding to a higher error rate

| Variable | *β* | Z-value | P-value |
|---|---|---|---|
| (Intercept) | −0.433 | −1.697 | <.1 |
| Correct suffix | −0.697 | −4.783 | <.001 |
| Adjacency | −4.830 | −13.267 | <.001 |
| Time: Afternoon | −0.037 | −0.270 | >.1 |
| Time: Evening | 0.345 | 2.553 | <.05 |
| Time: Night | −0.237 | −0.809 | >.1 |
| Followers | −0.184 | −7.096 | <.001 |
| Tweet length | −0.005 | −3.083 | <.01 |

From the variables used in previous experiments, *Correct suffix* was significant again. Unexpectedly, when the correct verb form had -<*dt*> as suffix, significantly *fewer* errors were made than when the correct suffix was -<*d*> . A significant effect of *Adjacency* was also found: Unexpectedly, when the verb form and the word determining its suffix were *not* adjacent, fewer errors were made.

With respect to the Twitter-specific variables, again an effect of *Time* was found. With *Morning* as baseline, significantly more errors were found in tweets

sent during the evening. Also, significantly more errors were made during the evening than during the afternoon ($\beta = 0.380$, $z = 2.870$, $p < .01$) and night ($\beta = 0.603$, $z = 2.060$, $p < .05$). Furthermore, the analysis again revealed an effect of *Followers*, with more followers corresponding to a lower error rate. Finally, the effect of *Tweet length* showed that fewer errors were made in longer tweets.

### 3.4 *Vinden*

The results of the analysis for *vinden* are summarized in Table 3. Again, *Relative Frequency* was not included in the analysis.

**Table 4.** Coefficients of Logistic regression on *Vinden* (*-d* vs. *-dt*), predicting incorrectness of target verb form, with higher $\beta$-values corresponding to a higher error rate

| Variable | $\beta$ | Z-value | P-value |
|---|---|---|---|
| (Intercept) | −3.216 | −10.276 | <.001 |
| Correct suffix | 3.200 | 14.074 | <.001 |
| Adjacency | −0.517 | −3.130 | <.01 |
| Time: Afternoon | 0.135 | 0.678 | >.1 |
| Time: Evening | 0.620 | 3.137 | <.01 |
| Time: Night | 0.388 | 0.887 | >.1 |
| Followers | −0.217 | −5.743 | <.001 |
| Tweet length | <0.001 | −0.017 | >.1 |

From the previously used variables, *Correct suffix* was significant again. When the correct suffix was -<*dt*> , significantly more errors were made than when the correct suffix was -<*d*> . A significant effect of *Adjacency* was found as well. In this analysis, the effect followed the expected pattern: Fewer errors were made when verb form and the word determining its suffix were adjacent than when they were separated.

From the Twitter-specific factors, again an effect of *Time* was found. Significantly more errors were found in tweets sent during the evening than in the morning, as well as in tweets sent during the evening, compared to the afternoon ($\beta = 0.485$, $z = 2.632$, $p < .01$). Furthermore, the analysis again revealed an effect of *Followers*, with more followers corresponding to a lower error rate. Finally, the effect of *Tweet length* was not significant.

## 4. Discussion

In this study, we investigated which factors play a role in spelling errors in homophonous inflected verb forms in Dutch everyday language behavior. To this purpose, we used a Twitter corpus containing 711,022 tweets. We extracted all homophonous verb forms ending in -<*d*> , -<*dt*> , and -<*t*> and classified them as correct/incorrect. Ultimately, 6.8% of all verb forms were found to be incorrectly spelled. In a logistic regression analysis predicting incorrectness, we found an effect of *Relative Frequency*, *Correct Suffix*, *Adjacency* of verb form and the word determining its suffix, *Time*, and *Number of Followers*.

When the intended verb form is less frequent than its homophone counterpart, more errors are made. This effect was previously found in many studies, starting from Assink (1985). The current study shows that the effect applies to spontaneously written text as well. This result suggests that also in everyday writing authors do not always apply the computational procedure (which should always give the correct output), but at least sometimes apply the retrieval procedure (which selects the most frequent form regardless of its correctness).

Furthermore, a preference for -<*d*> was found over both -<*t*> and -<*dt*> , supporting previous studies as well (e.g., Bosman 2005; Frisson & Sandra 2002; Sandra et al., 1999). This can be elucidated by combining two explanations. First, the preference for -<*d*> , rather than -<*t*> , could be due to overgeneralization (Neijt & Schreuder 2007). Through final devoicing, /*t*/ is sometimes spelled as -<*d*> in Dutch, while the reverse, /*d*/ written as -<*t*> , is systematically absent. This explanation is supported by Hanssen et al. (2015), who showed that Dutch first-graders initially have a -<*t*> -bias in their spelling, which turns into a -<*d*>-bias after they have learned the rule for final devoicing.

A second explanation which may account for the preference of -<*d*> over -<*dt*> is given by Ernestus and Mak (2005). They showed that people prefer analogy in the inflectional paradigm. As a consequence, when -<*dt*> is written in only one form of the inflectional paradigm (the third person singular, e.g. *beantwoordt* 'answers') while -<*d*> is used in all other forms (e.g., the first person singular *beantwoord*, the infinitive *beantwoorden*, the present participle *beantwoordend*, etc.), this leads to a preference for writing -<*d*> in the entire paradigm.

The overall preference for -<*d*> was found in all analyses, except for *worden*, where the -<*dt*>-form was preferred to the -<*d*> -form. Importantly, *wordt* is much more frequent (41101) than *word* (1209). This means that the preference for -<*d*> can be overruled by the difference in frequency between the two forms when this difference is large enough. This result suggests that when the frequencies of the homophones are close to each other, as for *vind* (3157) and *vindt* (3904), people prefer the -<*d*> -form, but when the -<*dt*>-form is much more frequent,

the preference shifts to that form. For further research, it might be interesting to investigate when the difference in frequency gets large enough to overrule the suffix preference effect.

A peculiar result in three of our four analyses was that the adjacency effect was either absent or reversed from the expected pattern. Based on earlier research starting from Assink (1985), we expected that fewer errors would occur when verb form and the word determining its suffix were adjacent, as the information needed to determine the suffix would still be salient in working memory. However, we found that people made *more* errors. This can be explained by the correlation between adjacency and relative frequency in our dataset: As the relative frequency increased (resulting in fewer errors), verb form and the word determining its suffix were more often separated. Apparently, the frequency effect prevents the adjacency effect to arise in situations less perfectly controlled than experiments. This is supported by the results of *vinden*, where adjacency *did* show the expected pattern. The frequencies of *vind* (3157) and *vindt* (3904) are very similar, suggesting that a facilitating effect of adjacency is only found when the frequency effect cannot arise.

Experimental studies are usually conducted during working hours. In contrast, we were able to compare data produced during the entire day and night. We found that relatively more errors are made during the evening and night than during daytime. An explanation could be that people get tired and less focused when the day proceeds. This is in line with Folkard (1975), who showed that both speed and accuracy of logical reasoning (associated with the use of working memory) decrease during the day. As demonstrated by Fayol et al. (1994), cognitive overload of working memory causes an increased spelling error rate as well. This overload may be reached earlier when people are tired.

Another factor assessed in the current study was the number of followers. Tweets written by users with more followers tend to contain fewer spelling errors. Several explanations can be given for this pattern. First, it can be a matter of prestige. Having many followers means that tweets are judged by many people, on the basis of content but also spelling. This could mean that twitterers are more cautious and perhaps more often double-check the spelling of their tweets when they have more followers. A second possibility is that people with many followers tend to have better language skills (including spelling) in general. When someone is talented in writing, people better like what they write, resulting in more followers.

We also investigated whether there was an effect of the length of the tweet. For Homophone Pair 1, more errors were made when the tweet was longer. This could either be caused by an increased complexity of the grammatical structure (e.g., Assink 1985) or by the simple fact that the longer a text is, the more likely it is that it contains more than one verb form. For Homophone Pair 2, no effect

of *Tweet length* was found. For *worden*, a reversed effect was found: Fewer errors were made in longer tweets. A possible explanation could be that in shorter tweets, *worden* is more often used as a copula, while in longer tweets it is more often used as an auxiliary verb, followed by a participle. The combinations of forms of *worden* plus the directly following word are probably more frequent when the verb forms function as auxiliary verb rather than as copula and may therefore be more often spelled correctly. This explanation also accounts for why the same was not found for *vinden* (no effect of *Tweet length*), as *vinden* always has the same function, independent of tweet length.

A small caveat has to be made regarding the number of errors found in this study. First, many tweets are sent from smartphones, meaning that autocorrection might already have corrected some of the errors. Secondly, we only focused on real homophones, discarding spelling errors leading to illegal spellings (pseudo-homophones). As pseudo-homophones are non-real words, no frequencies can be given for these forms, making comparison to real forms impossible. Thus, in reality, the total number of errors in spontaneous writing on Twitter is probably slightly higher than the found 6.8%. Nevertheless, this percentage is clearly lower than that in spelling experiments, such as the 25% found by Sandra (2010).

## 5. Conclusion

This study shows that spelling errors in Dutch verb forms are not only found in experimental settings, but also in tweets which reflect spontaneous, everyday writing. Our results support experimental findings by showing that the factors found in experiments largely apply to spontaneous written language production in a larger population as well. Furthermore, the analyses revealed effects of time and number of followers, which are unique for Twitter data. Overall, this study shows that Twitter is a very useful tool to supplement and critically reflect on laboratory research.

## Acknowledgements

We are grateful to Florian Kunneman for his help with accessing the Twitter database.

## References


Assink, Egbert M. H. 1985. "Assessing spelling strategies for the orthography of Dutch verbs." *British Journal of Psychology* 76 (3): 353–363. https://doi.org/10.1111/j.2044-8295.1985.tb01958.x

Baayen, R. Harald, Ton Dijkstra & Robert Schreuder. 1997. "Singulars and plurals in Dutch: Evidence for a parallel dual-route model." *Journal of Memory and Language* 37 (1): 94–117. https://doi.org/10.1006/jmla.1997.2509

Baayen, R. Harald, Richard Piepenbrock & Hedderick van Rijn. 1995. "The CELEX database." Nijmegen: Center for Lexical Information, Max Planck Institute for Psycholinguistics, CD-ROM.

Bosman, Anna M. T. 2005. "Development of rule-based verb spelling in Dutch students." *Written Language & Literacy* 8 (1): 1–18. https://doi.org/10.1075/wll.8.1.01bos

Ernestus, Mirjam & Willem M. Mak. 2005. "Analogical effects in reading Dutch verb forms." *Memory & Cognition* 33 (7): 1160–1173. https://doi.org/10.3758/BF03193220

Fayol, Michel, Pierre Largy & Patrick Lemaire. 1994. "Cognitive overload and orthographic errors: When cognitive overload enhances subject-verb agreement errors. A study in French written language." *The Quarterly Journal of Experimental Psychology Section A* 47 (2): 437–464. https://doi.org/10.1080/14640749408401119

Folkard, Simon. 1975. "Diurnal variation in logical reasoning." *British Journal of Psychology* 66 (1): 1–8. https://doi.org/10.1111/j.2044-8295.1975.tb01433.x

Frisson, Steven, & Dominiek Sandra. 2002. "Homophonic forms of regularly inflected verbs have their own orthographic representations: A developmental perspective on spelling errors." *Brain and Language* 81 (1–3): 545–554. https://doi.org/10.1006/brln.2001.2546

Hanssen, Esther, Robert Schreuder & Anneke Neijt. 2015. "From t-bias to d-bias in Dutch: Evidence from children's spelling and pronunciation." *Written Language & Literacy* 18 (1): 104–120. https://doi.org/10.1075/wll.18.1.05han

Hutto, C. J., Sarita Yardi, & Eric Gilbert. 2013. A longitudinal study of follow predictors on twitter. *Proceedings of the sigchi conference on human factors in computing systems*, 821–830). ACM. (http://comp.social.gatech.edu/papers/follow_chi13_final.pdf). https://doi.org/10.1145/2470654.2470771

Largy, Pierre, & Michel Fayol. 1996. "The homophone effect in written French: The case of verb-noun inflection errors." *Language and Cognitive Processes* 11 (3): 217–256. https://doi.org/10.1080/016909696387178

Laudanna, Alessandro & Cristina Burani. 1985. "Address mechanisms to decomposed lexical entries." *Linguistics* 23 (5): 775–792. https://doi.org/10.1515/ling.1985.23.5.

Neijt, Anneke & Robert Schreuder. 2007. "Asymmetrical phoneme-grapheme mapping of coronal plosives in Dutch." *Written Language & Literacy* 10 (2): 219–234.

Sandra, Dominiek. 2010. "Homophone dominance at the whole-word and sub-word levels: Spelling errors suggest full-form storage of regularly inflected verb forms." *Language and Speech* 53 (3): 405–444.

Sandra, Dominiek, & Michel Fayol. 2003. "Spelling errors with a view on the mental lexicon: Frequency and proximity effects in misspelling homophonous regular verb forms in Dutch and French." *Morphological structure in language processing* ed. by R. H. Baayen and R. Schreuder, 485–514. Berlin: Mouton de Gruyter.

Sandra, Dominiek, Steven Frisson & Frans Daems. 1999. "Why simple verb forms can be so difficult to spell: The influence of homophone frequency and distance in Dutch." *Brain and Language* 68 (1–2): 277–283.

Sandra, Dominiek, Steven Frisson & Frans Daems. 2004. "Still errors after all those years…: Limited attentional resources and homophone frequency account for spelling errors on silent verb suffixes in Dutch." *Written Language & Literacy* 7 (1): 61–77.

Sandra, Dominiek, & Lien van Abbenyen. 2009. "Frequency and analogical effects in the spelling of full-form and sublexical homophonous patterns by 12 year-old children." *The Mental Lexicon* 4 (2): 239–275.

Tjong Kim Sang, Erik & Antal van den Bosch. 2013. "Dealing with big data: The case of Twitter." *Computational Linguistics in the Netherlands Journal* 3: 121–134.

van der Veer, Neil, Steven Boekee & Oscar Peters. 2017. "Nationale Social Media Onderzoek 2017 [National Social Media Research 2017]." Accessed 27 March 2018. https://www.vonkfactor13.nl/wp-content/uploads/Newcom-Nationale-Social-Media-Onderzoek-2017.pdf.

Verhaert, Nina. 2016. "Rules Or Regularities? The Homophone Dominance Effect in Spelling and Reading Regular Dutch Verb Forms." PhD diss., Universiteit Antwerpen.

Verhaert, Nina, Ellen Danckaert & Dominiek Sandra. 2016. "The dual role of homophone dominance. Why homophone intrusions on regular verb forms so often go unnoticed." *The Mental Lexicon* 11 (1): 1–25.

## *Authors' addresses*

Tijn Schmitz
Centre for Language Studies
Radboud University
Erasmusplein 1
6500 HD Nijmegen
The Netherlands

t.p.a.schmitz@student.ru.nl

Robert Chamalaun
Centre for Language Studies
Radboud University
Erasmusplein 1
6500 HD Nijmegen
The Netherlands

robert.chamalaun@let.ru.nl

Mirjam Ernestus
Centre for Language Studies (CLS)
Radboud University
Erasmusplein 1
6500 HD Nijmegen
The Netherlands

m.ernestus@let.ru.nl