The impact of analogical effects and social factors on the spelling of partially homophonous verb forms in informal social media writing

Hanne Surkyn, Reinhild Vandekerckhove, & Dominiek Sandra

University of Antwerp

The impact of analogical effects and social factors on the spelling of partially homophonous verb forms in informal social media writing

Hanne Surkyn

University of Antwerp

Prinsstraat 13 (S.D.221)

2000 Antwerp, Belgium

hanne.surkyn@uantwerpen.be

Reinhild Vandekerckhove

University of Antwerp

Prinsstraat 13 (S.D.224)

2000 Antwerp, Belgium

reinhild.vandekerckhove@uantwerpen.be

Dominiek Sandra

University of Antwerp

Prinsstraat 13 (S.D.219)

2000 Antwerp, Belgium

dominiek.sandra@uantwerpen.be

The impact of analogical effects and social factors on the spelling of partially homophonous verb forms in informal social media writing

Abstract

The present study examines unintentional spelling errors on past participles produced by Flemish teenagers in private online writing. Previous psycholinguistic research on verb spelling errors in Dutch mainly focused on *identical* homophones (Bosman 2005; Frisson & Sandra 2002; Sandra et al. 1999). The present study, however, deals with past participles that are only partially homophonous with other forms in the inflectional paradigm and investigates whether the spelling of these verbs is affected by whole-word frequency, paradigmatic and bigram support for the correct spelling and the token frequency of the past participles' morphological family. The error rates reflect the effect of both paradigmatic and bigram support. Moreover, the unique database makes it possible to analyze the impact of three social factors (*Gender*, *Educational track* and *Age*). Our results reveal an effect on the error rates of all social variables. Finally, these social factors do not interact with paradigmatic and bigram support.

Keywords: spelling errors, past participles, homophones, analogical effect, morphological effect, bigram effect, gender, age, education, social media writing

Spelling errors on regularly inflected Dutch verb forms are persistent, even in experienced writers. Most of these errors are made on homophonous verb forms, which have the same pronunciation but a different spelling, e.g., *word* and *wordt* ('become' – 'becomes'). Over the past decades, researchers have extensively investigated spelling errors on these Dutch verb homophones (e.g., Bosman 2005; Chamalaun, Bosman, & Ernestus in press; Frisson & Sandra 2002; Sandra 2010; Sandra, Frisson, & Daems 1999; Schmitz, Chamalaun, & Ernestus 2018; Verhaert, Danckaert, & Sandra 2016). These psycholinguistic studies demonstrated that the spelling of verb homophones is affected by the relative frequency of the verb homophones: Most errors occur on the lower-frequency form. This effect is known as the homophone dominance effect and is explained as an interplay between an attentional process in working memory and an automatic process in long-term memory (Sandra et al., 1999). A conscious application of the spelling rules is required in order to spell the verb homophones correctly. This process has not become automatic, most probably because spellers of Dutch are rarely confronted with these verb forms (roughly, only in about 5 to 10% of all verb types and tokens), which cannot be spelled the way they are pronounced (Verhaert et al. 2016). This may cause a temporary overload of working memory resources, which, in turn, will increase the probability that an automatic process of lexical retrieval activates the higher-frequency spelling.

The present study differs from previous research in three respects. First, we examine spelling errors on *partially* homophonous verb forms, in particular on past participles that end in a <d> (pronounced as [t]) and are partially homophonous with other inflectional forms in the verbal paradigm ending in <t> (e.g., the past participle *genoemd* ('named'), which is partially homophonous with the simple present form

'noemt' ('names'), since apart from the prefix *ge-* both forms sound identical, i.e., /numt/). Note that verbs like *landen, '*to land*'* (past participle: *geland*)*,* with a stem-final <d>, do have identical homophones, i.e., the 1st and 3rd person simple present forms (*land* and *landt,* both pronounced as [lant]). The past participles of these verbs were not included in this study. To date, most studies focused on the effect of *identical* homophones, i.e., homophony at the whole-word level (see e.g., Chamalaun et al. in press; Frisson & Sandra 2002; Sandra et al. 1999; Schmitz et al. 2018; Surkyn et al. 2020; Verhaert et al. 2016). Contrary to the verb forms in some of these studies, the homophonous <t> spelling of the past participles in the present study does not render an existing form: The past participle *genoemd* has no variant *genoemt* in its inflectional paradigm, since the corresponding simple present form is *noemt*, without the prefix *ge-*. Hence, the spelling of this type of past participles cannot be influenced by an identical homophone. However, we hypothesize that it can nevertheless be affected by partial homophones, i.e., other forms in the verb's inflectional paradigm that are homophonous with the letter string in the past participle following the *ge-* prefix. In other words, the simple present *noemt* ('names') may trigger an incorrect <t> spelling in *genoemd*. By contrast, the simple past *noemde* ('named'), which is pronounced as [numdə], may be an extra trigger for a <d> ending when spelling the past participle *genoemd,* thus creating an analogical effect that counteracts the effect of the partial homophone *noemt*. If so, the frequencies of these inflectionally related forms, whether partially homophonous (e.g., *noemt*) or offering <d> support (e.g., *noemde*), could modulate analogical effects.

A perception study conducted by Ernestus and Mak (2005) indeed demonstrated the importance of analogical effects in reading Dutch verb forms. Their study revealed

that an incorrectly spelled verb form (an incorrect <d> spelling) caused a smaller reading delay when other inflected forms of the same verb supported this spelling. In the present study, we will examine whether such supportive effects can also be observed in the (spelling) production of these Dutch past participles.

Second, in addition to the psycholinguistic focus on frequency effects in the production of verb spelling errors, we include a sociolinguistic dimension by analyzing the impact of the social variables *gender*, *age* and *education*, which gives our research an interdisciplinary character (see Surkyn et al. (2020) for a comparable approach for identical verb homophones). By combining a sociolinguistic and a psycholinguistic perspective, we are also able to test whether social factors interact with the mental processes that determine the error risk. We follow the same rationale as in a recent study (Surkyn et al. 2020), where we showed, for two types of verb homophones, that the three social factors mentioned above did not interact with the effect of homophone dominance. This constituted independent proof for the model of Sandra and colleagues (1999). Indeed, the model predicts that different social groups differ only at the level of attentional processes in working memory, not at the level of unconscious processes, i.e., the automatic retrieval of orthographic representations. Different groups are thought to be more or less concerned about spelling correctness or to be faster or slower in applying the spelling rule, which will affect their error risk. However, this will not affect the automatic processes of retrieving orthographic representations. In other words, it will not affect which orthographic representation tends to be (erroneously) retrieved from long-term memory, i.e., the higher-frequency homophone in our previous study. Following the same rationale, we expect no interaction between the social and psycholinguistic variables in the errors on partially homophonous verb forms.

Finally, by exploring private chat conversations produced on WhatsApp and Facebook Messenger, we examine verb spelling errors in real-life data. In this way, we test the ecological validity of previous psycholinguistic experiments. Recent studies demonstrated that earlier experimental findings indeed generalize to contexts outside the lab: The homophone dominance effect was observed in tweets by Schmitz et al. (2018) and in private text messages on WhatsApp and Facebook Messenger in our previous studies (Surkyn et al. 2019; Surkyn et al. 2020). We will now test whether the theoretical framework behind this effect can be generalized to verb forms that are not fully homophonous. In the absence of identical homophones, no measure of homophone dominance could be calculated. However, for these partial homophones, another relationship might determine spellers' tendency to write the correct <d> spelling. More particularly, we will examine whether the error risk is determined by (a) a frequency ratio that reflects the support for a <d> spelling versus a <t> spelling in the verb's inflectional paradigm and (b) a frequency ratio reflecting <d> support versus <t> support in the final bigrams of (all) Dutch verb forms, i.e., a potential determinant at a much lower level. By using non-elicited spontaneous online writing data, our results might further corroborate the ecological validity of the interpretational frame derived from the experiments.

It is important to emphasize that the errors are unlikely to be intentional errors that are induced by the social media context, even though informal online writing tends to diverge from formal writing in several respects. In informal computer-mediated communication (CMC), people for instance tend to ignore standard writing conventions by intentionally manipulating spelling and typography (e.g., by means of letter repetitions as in *suuuuper*) or by applying all kinds of strategies to shorten words and

utterances (e.g., *np*, '*no problem*'). Another typical marker of the genre is the use of emoji. These and many other prototypical features of informal CMC can be described in terms of three pragmatic maxims (e.g., Androutsopoulos 2011): the orality maxim (i.e., mimic spoken language, e.g., vernacular forms), the economy (or speed) maxim (i.e., be as brief as possible), and the maxim of expressive compensation (i.e., compensate for the lack of the non-verbal expressive cues that are part of face-to-face communication). Importantly, the present paper does not focus on these deliberate and genre-specific deviations from standard writing, but on unintentional spelling violations. The spelling errors on Dutch verb forms under consideration are likely unintentional: They are no markers of the genre and, in view of their negative perception (these errors are generally stigmatized), there is no prestige to be gained by producing them.

The spelling of Dutch past participles

Dutch past participles generally consist of the following elements: a prefix (mostly *ge-*) + the verb's stem + a suffix.[1] The suffix is spelled as <d> or <t>, depending on a phonological property of the stem-final sound, i.e., the infinitive minus the infinitival suffix *-en*, e.g., *werk(en)* ('to work'). If this final sound is a voiceless consonant, the past participle is spelled with a <t> ending. If the verb stem ends in a voiced sound, the <d> spelling is correct. The same morphophonological conditions apply in the formation of the simple past, where the singular and plural suffixes *-de/-den* and *-te/-ten* are used after voiced and voiceless stem-final consonants, respectively. This makes the simple past an analogical model for the past participle. However, in order to determine

[1] The prefix of separable compound verbs with emphasis on the first part is placed between that first part and the basic verb, e.g., *aan-ge-leerd*.

whether a consonant is voiceless or voiced, an explicit comparison with the simple past is not required. In order to avoid metalinguistic analyses, children are taught they can use mnemonics such as *'t kofschip* (a type of sailing ship) or *'t fokschaap* (a breeding sheep). These words contain all voiceless Dutch consonants that can occur at the end of a verb stem.

We focus on partially homophonous past participles of regular verbs that are spelled with a <d> ending, although a final [t] sound is heard due to final devoicing (e.g., *ge-noem-d*, 'named'). Thus, in order to spell these past participles correctly, spellers cannot rely on the phonological principle ('spell what you hear', which would result in e.g., **genoemt*). Instead, they have to take into account the morphological analogy principle that determines the spelling of verb forms in the Dutch language: The past participle *genoem-d* has to be spelled with a <d> in accordance with the simple past form *noem-de* which also has a suffix including a <d>. As mentioned above, spellers can also rely on mnemonics such as *'t kofschip* (see above) and conclude that the final sound of the verb's stem is not included in this word, which means that the <d> spelling is the correct one.

In spite of the informal online writing context, it seems unlikely that the incorrect, phonetic <t> spellings are the result of intentional spelling manipulations. While people often mimic spoken language in social media writing (see the orality maxim, Androutsopoulos 2011), for example by spelling words phonetically, this 'write as you speak' principle is seldom applied systematically. An accurate and consistent representation of speech simply is not aimed at in informal CMC (see Vandekerckhove & Nobels 2010: 669). With respect to the spelling of Dutch verb forms, our previous study (Surkyn et al. 2020) on verb spelling errors in informal online writing showed that

adolescents strongly preferred the morphographic, i.e., non-phonetic, spelling of the verb stem when writing (homophonous) verb forms. In the analysis of stem final <d> verbs (e.g., *vinden*, 'to find') only 14 phonetic spellings (i.e., <t> spellings) on 5,804 verb forms were observed (e.g., *vint* instead of *vind*/*vindt*, 'find/finds') (Surkyn et al. 2020: 459). This result suggests that adolescents do not apply the orality maxim when spelling Dutch verb forms. It also shows that the effect of homophone dominance reflects problems with the morphographic spelling of the suffix, not with the spelling of the stem, i.e., two different morpheme types at which the morphological principle in Dutch spelling is applied.[2] Based on these results, it seems safe to assume that the (incorrect) <t> spellings of the past participles in this study, like the spelling errors in our previous study, are unconscious spelling errors (affected by social and psycholinguistic factors) rather than intentional spelling manipulations.

1. Research Questions and Hypotheses

In the present paper, we study the effect of social and psycholinguistic factors on the production of spelling errors on partially homophonous past participles in adolescents' online chat conversations.

1.1. Psycholinguistic variables

First of all, we examine the effect of the frequency of the past participle itself (i.e., its whole-word frequency) and the cumulative frequency of the morphological family

[2] The most likely explanation is that the morphological principle is applied extremely frequently at the stem level (both type-wise – across lexical categories – and token-wise) but very infrequently, i.e., in less than 1 out of 10 verb forms, at the suffix level (both type-wise and token-wise) (see Verhaert et al. 2016).

consisting of the past participle of the verb's stem and all its compound past participles (henceforth, *PP family frequency*; e.g., for *gestuurd,* '(has) steered'*,* and *opgestuurd,* '(has) sent', this morphological family would comprise these two forms and all other past participles containing *gestuurd* as their rightmost constituent, such as *ingestuurd,* 'submitted' and *uitgestuurd,* 'sent out'). We distinguished between these two frequency measures because the probability that one of the members in this PP family is spelled correctly may not only be determined by the frequency of the simple past participle but also by the frequencies of all compounds in which it occurs. We hypothesize that the frequency of the past participle itself or of the whole compound family will determine the error rates.

In addition, we investigate whether the spelling of the past participles is affected by the frequency of other verb forms in the verbal paradigm, whether partially homophonous ones (e.g., *noemt* for *genoemd*) or not, i.e., forms supporting the <d> spelling (e.g., *noemde* for *genoemd*). Since the past participles under investigation require a <d> spelling, we hypothesize that a greater support for the <d> form in the verbs' inflectional paradigm will cause fewer errors (see also the perceptual research by Ernestus and Mak (2005) referred to above).

Finally, we examine the impact of the frequency of word-final bigrams ending in <d> and <t> on the past participles' spelling. These bigrams include the stem-final letter and the *-d* or *-t* suffix of the past participles (e.g., <md> or <mt> for *genoemd*/*genoemt*, 'named'). As explained above, the stem-final phoneme determines the spelling of the past participles' suffix, only voiced consonants giving rise to the <d> spelling. This results in a high frequency (both type- and token-wise) of most bigrams whose first

letter reflects a voiced stem-final phoneme.[3] This does not mean that each bigram ending in <d> occurs more frequently than its corresponding bigram ending in <t>. It does mean that spellers are highly familiar with these <d> bigrams (as the result of the morphographic spelling rule for the past participle) and that the degree of <d> support from this variable may affect their tendency to spell <d>, i.e., the correct spelling. We retrieved all verb forms from SUBTLEX-NL whose final bigram ended in <d> or <t> and calculated, for each initial letter of these bigrams, its type and token frequencies when followed by <d> and <t>. SUBTLEX-NL is a database of Dutch word frequencies based on more than 40 million words from television and film subtitles (Keuleers, Brysbaert, & New 2010). The correlation between the type and token measures across this set of voiced stem-final phonemes was high: .65 (measure = proportion <d> spellings) and .60 (measure = log odds of <d> spellings). In both counts, the bigrams <ld>, <rd> and <wd> were among the four highest-ranking bigrams. Hence, both type and token frequencies create familiar bigrams ending in <d> in verb-final position. The variability on these variables for <d> support allow sufficient room for assessing an effect of *Bigram <d>-support.*

As the past tense is the analogical model for the spelling of the past participle (see above), this would indicate a morphologically conditioned effect of bigram frequency. In contrast, if bigram-based support for the <d> spelling also depends on the distribution of the final bigrams in nouns and adjectives, the effect would be purely

[3] The phonemes /v/ and /z/ surface in these past participles as respectively <f> (e.g., *fuiven-gefuifd,* 'party'- 'partied') and <s> (e.g. *verhuizen-verhuisd,* 'relocate'-'relocated'), because <vd> and <zd> are orthographically illegal bigrams. As <f> and <s> are often followed by <t> in verb forms, the proportion of <d> bigrams is the smallest for these phonemes, both type-wise and token-wise.

orthographic. As with the above variable, we assume that a greater <d> support from the stem-final letter will lead to fewer errors. Note that this bigram frequency factor, too, would be a measure of <d> support, but at the sublexical level.

As the final bigrams ending in <d> are homophonous with the final bigrams ending in <t>, due to final devoicing (for example, <rd> and <rt> are both pronounced as [rt]), an effect of <d> support from the bigrams would suggest an effect of homophony at the sublexical level. Previous research of Sandra (2010) and Sandra & Van Abbenyen (2009) already revealed an effect of homophony at the sublexical level, even though this effect was situated at a higher level than bigrams, more particularly, orthographic patterns straddling the stem-suffix boundary of regular past tenses. In a corpus study conducted by Sandra (2010) more spelling errors were made on the past tense of verbs whose stem ends in the <st> cluster (e.g., *taste* instead of *tastte*, 'groped') than of verbs whose stem ends in the <cht> cluster (e.g., *wachte* instead of *wachtte*, 'waited'), as *ste* (with a single <t> instead of a double <t>) occurs proportionally much more often in regular past tense forms in the *s* group dan *chte* in the *ch* group (Sandra 2010, Corpus Study 1: 430). Hence, the spelling of regular past tense forms was affected by the relative frequencies of sublexical homophonic spelling clusters.

An experimental study conducted by Sandra & Van Abbenyen (2009) revealed a similar effect of sublexical homophony. In a spelling experiment (among 12-year-old children) three types of past tenses with stem final <d> and suffix *-de* had to be filled out (e.g., *raadde*, 'guessed', *leidde*, 'led', *meldde*, 'reported'), differing in the type of orthographic cluster preceding the stem <d> (Sandra & Van Abbenyen 2009: 239). However, only one verb type was characterized by the presence of sublexical homophony. For example, the sublexical orthographic pattern <ldde> at the end of the

form *meldde* (‘reported’) was homophonous with the spelling sequence <lde> at the end of forms like *belde* (‘called’) or *telde* (‘counted’) (Sandra & Van Abbenyen 2009: 253). Results showed that significantly more errors were made on verbs with sublexical homophony in their past tense system than on verbs without such homophony at the sublexical level. Additional analyses demonstrated a higher occurrence frequency (both type-wise and token-wise) of past tenses with the orthographic end cluster with a single <d> (e.g., <lde>, <rde>, <nde>) than past tenses with the final clusters with a double <d> (e.g., <ldd>, <rdde>, <ndde>). Hence, Sandra & Van Abbenyen (2009) conclude that the errors on these past tense forms (spelled with a double <d>) are the result of sublexical competition from the homophonous spelling pattern (in the past tense) with a single <d>. Similarly, in the present study we will check whether the homophonous bigrams ending in <t> (e.g., <rt>) and <d> (e.g., <rd>) affect the spelling of past participles ending in <d> (e.g., <rd> in *geleerd*, ‘learned’).

1.2. Social variables

Furthermore, we will examine whether the social variables *Gender*, *Educational track* and *Age* affect the error rates. Previous research on the effect of *gender* on the spelling of identical verb homophones by teenagers in informal online writing demonstrated that girls make fewer verb spelling errors than boys, probably as a result of a difference in error awareness and norm sensitivity (see Surkyn et al. 2019; Surkyn et al. 2020). In sociolinguistic studies, women have been reported to “conform more closely than men to sociolinguistic norms that are overtly prescribed” (Labov 2001: 293) and to avoid stigmatized forms (e.g., Tagliamonte 2011: 32). Spelling errors on Dutch verb forms are indeed highly stigmatized due to the clear-cut rules. We therefore expect girls to

allocate more attention to the correct spelling and to avoid these stigmatized errors more than boys, all the more so since this gender effect has already been found for identical verb homophones and there is no ground to assume that the partially homophonous verb forms would trigger other social effects than these homophones. In Surkyn et al. (2019) we suggested that the effect of gender could also be an effect of message length in disguise. Previous research has demonstrated that boys tend to write shorter messages than girls (Hilte 2019: 159), which might indicate that boys focus more strongly on speed and hence produce more errors. Consequently, we decided to include the factor *Post length* in our analyses.

With respect to the social variable *Educational track* (see 2.2.2), adolescents in the more theory-oriented tracks of (Flemish) secondary education have been observed to produce fewer verb spelling errors than their peers in the more technical or practice-oriented educational tracks (Surkyn et al. 2020; Van den Bergh, Van Es, & Spijker 2011; Van der Horst, Van den Bergh, & Evers-Vermeul 2012; in error detection, see Verhaert et al. 2016). Students in the more theory-oriented track are likely to have a better knowledge of verb spelling rules and better metalinguistic skills, as their official curricula tend to attach more importance to (reflection on) spelling issues than the curricula for more practice-oriented tracks, where proofreading and a proper use of resources and correction tools (such as online dictionaries) are deemed more important than training the explicit knowledge of spelling and grammar rules (VVKSO 2006, 2014). In practice-oriented tracks, Dutch language teaching is not even a separate course. Instead, it is part of a larger package of general subjects which is called *Project Algemene Vakken* or *PAV* ('project general subjects') and also includes mathematics, geography, history and biology. Its focus is on functional language skills, not (at all) on

rule knowledge and grammatical insight. Recently, Chamalaun, Bosman and Ernestus (in press) demonstrated that many verb spelling errors result from the inability to determine the grammatical role of the verb form, especially in the more practice-oriented tracks. Hence, we predict higher error rates in the more technical or practice-oriented tracks.

Finally, for the factor *Age,* we make a distinction between older adolescents (or young adults) and younger adolescents (see 2.2.2) and expect older adolescents to make fewer spelling errors than younger ones. The latter are likely to attach less importance to writing and spelling conventions than the older adolescents, since non-conformist linguistic behavior tends to be at its highest in early adolescence and tends to decrease as teenagers approach adulthood. This phenomenon is often referred to as the 'adolescent peak' (Coates 1993: 94; Holmes 1992: 184). An alternative account is that older adolescents are better in determining the grammatical role of the verb form (see Chamalaun et al. in press), which might result in fewer errors on their part.

1.3. Interaction between psycholinguistic and social variables

Finally, the account proposed by Sandra et al. (1999) leads to the prediction that social factors will not interact with the frequency variables. If an effect of the social factors can be observed, this will be at the level of conscious processing, i.e., when determining the suffix spelling in working memory. This can either be related to the willingness to consciously reduce the number of errors or to the ability to engage sufficiently fast in applying the spelling rule. In contrast, the social variables should have no influence on the unconscious frequency-driven retrieval of stored orthographic forms. This

prediction was confirmed in Surkyn et al. (2020), who found no interaction between homophone dominance in identical homophones and gender, age, and educational track. Hence, we assume that certain social groups produce more errors than others, but that all social groups will undergo the same effects of psycholinguistic frequency factors.

2. Methodology

2.1. Corpus

The corpus contains informal, private online chat conversations written by Flemish adolescents aged 15 to 20 years old. All 403,491 posts or 2,360,117 tokens[4] were produced via Facebook Messenger or WhatsApp and nearly all of them were written in 2015 and 2016.[5] The teenagers were secondary school students at the time. Most of them lived in the province of Antwerp, one of the (central) provinces of Flanders, i.e., Dutch-speaking Belgium. The collection of the data was supported by teachers and school principals. The same dataset was used by Surkyn et al. (2020), which makes a direct comparison between the studies possible.

The distribution of the data over the three social variables is presented in Table 1. We signal that girls and students in technical education are overrepresented in terms of posts, tokens and target forms. This imbalance is due to the procedure for data collection. The adolescents were requested to voluntarily donate their chat posts: They

[4] A token is a visual unit used in isolation or separated by a blank space from the preceding unit (e.g., a word or an emoji that does not stick to a preceding word).

[5] The original corpus, as described in Hilte (2019: 17), also included messages of 13-and 14-year-olds. Since we have much less data (and fewer target forms) at our disposal from these youngest teenagers than from the older ones, their messages were excluded from our analyses.

were free to send us as much conversations as they wanted. Consequently, some provided us with much more data than others.[6] Nevertheless, even for the smallest group (professional education students), the corpus contains almost 1,500 past participles that are relevant for the present study.

(INSERT TABLE 1 ABOUT HERE)

2.2. Research Variables and Their Operationalization

2.2.1. Dependent Variable

Spelling performance on Dutch partially homophonous past participles that end in a <d> is the dependent variable in the present study. We manually encoded the spelling of all relevant past participles in the corpus as 'correct' or 'incorrect'. Only the (correct) <d> and (incorrect) <t> endings were taken into account. This means that verb forms with a <d> ending were categorized as 'correct' even if they contained typos (e.g., *gekluisterd* instead of *geluisterd*, 'listened') or other spelling errors (e.g., *geinformeerd* instead of *geïnformeerd*, 'informed').[7] Since we focus on *partially* homophonous past participles, past participles such as *gebeurd* ('happened'), which are fully homophonous with the second and third person singular simple present (*gebeurt*, 'happens', the prefix *ge-*

[6] All chat data were anonymized and the procedure was given clearance by the ethical committee for the Social Sciences and Humanities of the Antwerp University, according to the GDPR guidelines.

[7] In 32 cases, the past participle was abbreviated by omission of 1 or more vowels (e.g., *gezgd* for *gezegd*, 'said'). These abbreviated past participles were included and treated in the same way as their unabbreviated equivalent: Verb forms with a final <t> spelling were encoded as 'incorrect', verb forms ending in <d> were encoded as 'correct'.

already being part of the stem), were not part of the dataset. Furthermore, we excluded past participles that function as an attributive adjective (e.g., *een gemotiveerd team*, 'a motivated team')[8] as well as verbs borrowed from English (e.g., *gegamed*, 'gamed'; *getagd*, 'tagged'), as the spelling of these morphologically integrated English verbs poses additional challenges and causes a lot of confusion.

2.2.2. Independent Variables

Psycholinguistic variables

Recall that the correct ending for all of the selected past participles is <d>. The participles are partially homophonous with simple present forms ending in <t>, but other verbal forms in the inflectional paradigm (e.g., the singular and plural simple past forms) actually do support the <d> ending. Therefore, so-called *Paradigmatic <d> support* is the first independent variable in the present study. This variable was operationalized as (the logarithm of) the ratio between the summed frequencies of the <d> forms over the summed frequencies of the <t> forms in the verbs' inflectional paradigm. These frequencies were extracted from SUBTLEX-NL (Keuleers, Brysbaert, & New 2010). This frequency variable is thus a continuous one, with higher values corresponding to a greater <d> support. The frequencies of the <d> forms were

[8] These attributively used past participles were excluded from the analyses because we wanted to maintain equivalence with previous research (Surkyn et al. 2020) in which we studied past participles in verb phrases, i.e., past participles accompanied by an (explicit or implicit) auxiliary verb. Moreover, it is (a priori) not clear whether a past participle behaves identically when it has an attributive function versus when it is part of a verb phrase. In the present study, our focus is exclusively on verb forms (as in our previous studies).

obtained by summing over the frequency of the past participle itself (e.g., *geleerd,* '(has) learned'), the simple past forms (singular and plural, e.g., *leerde, leerden,* 'learned'), the (inflected) present participle (*lerend, lerende,* 'learning') and the (inflected) attributively used past participle (*geleerd, geleerde,* 'learned'). These verb forms are all inflectional variants of the verb with a <d> spelling. The frequencies of the <t> forms were obtained by summing over the frequencies of the 2nd and 3rd person singular simple present.[9] Table 2 gives an overview of the computation of the frequencies of the <d> and <t> forms.

(INSERT TABLE 2 ABOUT HERE)

For each relevant <d> or <t> form, we calculated the cumulative frequency of all verb forms in this grammatical form sharing the base verb. Thus, when the target form was a

[9] The past participles *gehad* ('had') and *gekund* ('could') were removed from our data since the 2nd and 3rd person singular simple present of these verbs (*heeft*, 'has' and *kan*, 'can') are not homophonous on a morphological level with the past participle. Recall that we also excluded all past participles (n = 320) of verbs with a stem-final <d> as some of these verbs have past participles with identical homophones in the singular present tense (e.g., *verbranden,* 'to burn', *verbrand*, '(has) burnt', *verbrand*, '(I) burn', *verbrandt*, '(you/he/she) burns', all pronounced as [vərbrant]). Other verbs with a stem-final <d> have identical homophones in the simple present singular, i.e., the 1st person and the 2nd and 3rd person (e.g., *landen,* 'to land', *land,* '(I) land', *landt*, '(you/he/she) lands'). Moreover, in some cases, the latter type of 1st person form is also homophonic and homographic with a noun (e.g., *(het) land,* '(the) land'). The past participles of these two verb types were excluded because, in contrast to our target forms, they have identical homophones in their inflectional paradigm and receive extra <d> support of the 1st person singular simple present.

compound verb (e.g., *aanleren*, 'to teach'), the frequencies of the <d> and <t> forms of the compound verb itself were added to the frequencies of the same grammatical forms of the base verb (e.g., *leren,* 'to learn') and of all other compound verbs of this verb (e.g., *afleren,* 'to unlearn'). For instance, the frequency of the <t> form for *aanleert* ('teaches') was determined by summing over the frequencies of *aanleert, leert, afleert* and all other compound verbs with *leren* ('to learn') as their final constituent. Vice versa, the frequencies of the word forms of compound verbs (*aanleren*) were added to the frequencies of the same grammatical forms of the base verb (*leren*) and all compounds with this verb in final position.

The second independent variable is *Bigram <d> support*, i.e., the <d> support in word-final bigrams. We operationalized this variable in the same way as our variable *Paradigmatic <d> support*: the logarithm of the ratio between the summed frequencies of the bigrams ending in <d> (e.g., <rd>, <gd>, <md>) over the summed frequencies of the bigrams ending in <t> (e.g., <rt>, <gt>, <mt>). In a first operationalization of the variable, we restricted the word-final bigrams to Dutch verb forms (e.g., *word*, 'become'; *vermoord*, 'killed'; and *hoort*, 'hears'; *probeert*, 'tries', for the frequency of <rt> and <rd>, respectively). In a second operationalization of this variable, we focused on word-final bigrams in Dutch words from lexical categories other than verbs (e.g., the adjective *hard* ('hard') and the nouns *moord* ('murder'), *woord* ('word') and *soort* ('kind'), *hart* ('heart'), *buurt* ('neighborhood') for the frequencies of <rd> and <rt>). In both cases, frequencies were extracted from SUBTLEX-NL (Keuleers, Brysbaert, & New 2010). Similar to *Paradigmatic <d> support*, this new frequency variable is a continuous variable, with higher values corresponding to a greater <d> support for the stem-final grapheme.

Two additional predictors were the frequency of the past participle itself, henceforth *Whole-word frequency*, and the frequency of the PP family, i.e., the cumulative frequency of all past participles containing the past participle of the base verb (e.g., *geleerd, aangeleerd, afgeleerd, bijgeleerd*, …).[10]

In view of the potential interference between *Gender* and *Post length* (the latter possibly being related to writing speed, see above), we included *Post length* as a final independent variable. This variable was calculated as the number of tokens (i.e., visual units, see above) per post.

Social variables

The research design includes three social variables. The first social variable is *Gender,* which was operationalized as a binary variable, since students identified themselves as

[10] We use the terms 'inflectional paradigm' and 'PP family' rather than the term 'morphological family' (see e.g., Schreuder & Baayen 1997) in order to avoid terminological confusion. In contrast to a typical morphological family, as intended by Schreuder and Baayen, the past participles under investigation have no nominal or adjectival derivations and/or compounds. Some of them do have morphologically more complex (or less complex) past participles that form a small morphological family (e.g., *gevraagd,* 'asked'; *afgevraagd,* 'wondered', *opgevraagd,* 'requested'). We measure the impact of this family with our factor 'PP family'. The verb's inflectional variants whose suffix contains a <d> or <t> are subsumed under our term 'inflectional paradigm'; they define our measure of *Paradigmatic <d> support*. Note that another verb type in Dutch, which we have also studied in previous studies (Surkyn et al. 2019, 2020), i.e., stem-final <d> verbs like *antwoorden* ('to answer'), do have derived and/or compound nouns and adjectives in their morphological family (e.g., the derived noun *antwoord,* 'answer', the derived adjective *beantwoordbaar, 'answerable',* and the compound *antwoordapparaat,* 'answering machine'). However, these verbs were not included in our study.

boys or girls. Hence, it can be assumed that *Gender* largely (but not necessarily completely) coincides with biological sex.

Educational track is the second social variable. We distinguished between adolescents on the basis of their educational profile. All participants were enrolled in one of the three main educational tracks of Flemish secondary education, i.e., in *aso*, *tso* or *bso*. *Aso* (*algemeen secundair onderwijs*) represents general secondary education (GE), which has a strong theoretical orientation and prepares for higher (university) education. *Bso* (*beroepssecundair onderwijs*) refers to professional secondary education (PE). This track prepares for direct access to the job market (mainly for manual labor). *Tso* (*technisch secundair onderwijs*) is technical secondary education (TE), holding an intermediate position between *aso* and *bso* as it has a practical as well as a theoretical orientation. These students usually proceed to higher education, though less often to university than their aso peers.

The third social variable relates to the age of the adolescents. We made a distinction between younger (15-and 16-year-old) and older (17-20-year-old) adolescents. This subdivision was based on the grade structure in Flemish secondary education: The 15- and 16-year-olds are the second graders of secondary education, the 17- to 20-year-olds are the third graders (The first graders, not included in the present research design, are the 13- and 14-year-olds). This division into grades is made in all educational tracks, and separate curricula are formulated for each grade. Henceforth, we will refer to the younger group as *Grade 2* and to the older group as *Grade 3*. Thus, Age is not treated as a continuous but as a categorical variable. This decision is further supported by previous sociolinguistic findings that teenagers' linguistic nonconformity, which manifests itself, for instance, through the use of adolescent slang or non-standard

language, does not evolve linearly as they get older. Instead, studies suggest that it increases until the age of 15 or 16 (mid-puberty) and then decreases again (i.e., the 'adolescent peak', see e.g., Holmes 1992: 184).

2.3. Data processing

In an automated prefiltering process all potentially relevant verb forms were extracted from the corpus by selecting all tokens containing <ge> and ending in <d> or <t>. As a result, there were no false negatives and maximum coverage in terms of recall was achieved. Next, in a post processing phase all false positives were removed manually. Most of these false positives concerned tokens that, without context, could be misinterpreted as past participles. For example, the token *gemotiveerd* ('motivated') can either function as an adjective (*Een gemotiveerd team,* 'A motivated team') or as a past participle (*Hij heeft haar gemotiveerd*, 'He has motivated her').[11] Additionally, we excluded standard messages generated by Facebook Messenger and WhatsApp (e.g., *bestand bijgevoegd*, 'file attached'). The final dataset consists of 6,551 target forms (distributed across 685 chatters and 382 verbs). Each target form was not only encoded for correctness but also for the social and psycholinguistic independent factors mentioned above.

[11] Note that there is no inconsistency in removing adjectival uses of past participles from the error analyses while including them in our measure of <d> support. We restricted our attention to spelling errors on the verbal uses of past participles that end in <d>, but this obviously does not imply that the spelling of the adjectival uses does not contribute to the overall <d> support and thus to writers' tendency to spell a <d>.

The data are unevenly distributed across verbs: As Figure 1 shows, most past participles occur only once or twice in our corpus. However, a handful of past participles are highly frequent (e.g., *gezegd* ('said') occurs 1,544 times, *gestuurd* ('sent') 854 and *gevraagd* ('asked') 522 times).[12] We previously demonstrated that this unbalanced nature of the dataset requires extra caution since one single high frequency verb can distort the results (see Surkyn et al. 2020). We provide more information on the way we dealt with this issue below.

(INSERT FIGURE 1 ABOUT HERE)

A detailed analysis of the data (and the outliers in particular) revealed that the past participle *gezegd*, '(has) said', which provides no less than 23.57% of all observations, is not only an outlier in terms of frequency.[13] It also stands out in terms of the error rate: While the overall error rate is 14.46%, no less than 26.81% of all forms of *gezegd* were spelled with an incorrect <t> ending.[14] Omission of *gezegd* from the database results in a general error rate of 10.65%. In order to avoid a complete distortion of the results by this one atypical, highly frequent past participle, we removed *gezegd* from our analyses.[15] The omission of *gezegd* led to a final corpus of 4,978 tokens.

[12] This is predicted by Zipf's law (Zipf 1949), which states that there are huge frequency differences among the words in a language.

[13] There were no such outliers from the perspective of chatters. The most frequent chatter produced only 2.72% of all target forms.

[14] A possible explanation of this high error rate is presented in the General Discussion.

[15] *Afgezegd* ('cancelled'), *opgezegd* ('cancelled') and *toegezegd* ('pledged'), i.e., compound past participles of *gezegd*, were also excluded from the analyses.

As noted above, besides *gezegd*, there are other past participles with a high occurrence frequency. However, the only other two high-frequency verbs (and their compounds) in the corpus that also strongly contributed to the overall error rate, i.e., *gestuurd,* 'steered', and *gevraagd,* 'asked', were not extreme outliers like *gezegd.* The latter past participle accounted for more tokens than these two forms together (i.e., 1,385). Moreover, their error rates, though high, were considerably lower than those for *gezegd:* 10.39% and 15.72%, respectively. Their extra removal would have resulted in an extra data reduction of 21.14%, such that the analyses would have been performed on only 55% of all data. Moreover, a preliminary inspection of the data revealed a different error pattern for *gezegd* than for the majority of verbs. Using the interquartile method for the removal of outliers could not solve the problem of skewed data either, as only 18.59% of the dataset would be left (1,218 observations). Nevertheless, since it is imperative to address this property of our database, we used a resampling approach.

We drew 1,000 random samples from all (4,978 remaining tokens of) past participles that occurred with a high frequency in the corpus. The cutoff for a high-frequency form, i.e., the maximum token frequency allowed in the sample, was determined on the basis of the distribution of the occurrence frequencies of the past participles in our corpus. As Figure 1 illustrates, an occurrence frequency of 100 tokens is a clear cut-off point: It marks a breaking point in the curve between the frequencies of *geprobeerd,* 'tried', and *gespeeld*, 'played' (86 and 135, respectively). In each run of the 1,000 simulations a random sample of 100 tokens was drawn from every verb form that had more tokens than this cutoff-value.[16] This procedure resulted in 3,093 tokens in

[16] All tokens of past participles that were less frequent than the cutoff were included in each run of the simulation.

each run. This way, we reduced the weight of some verbs so that they had less impact on the results. For example, the relative contribution of a high-frequency form such as *gestuurd* ('sent') dropped from 13.04% to 3.10%. Importantly, the random selection had a negligible effect on the number of chatters that were omitted: No chatter was systematically left out from all 1,000 runs and 478 out of 629 chatters (i.e., 75.99%) were included in all runs. In the subset of the 151 chatters that were dropped in 1 or more runs (range: 2 – 897), chatters were excluded 585 times on average (median = 643). Moreover, the random sampling procedure had no impact on the distribution of the data over the social variables. The mean number of tokens in the simulations accurately reflected the distribution of tokens in the database (i.e., without *gezegd* and its PP family): for *Educational track* (GE: 32% vs. 30%, TE: 47% vs. 48%, PE: 21% vs. 22% in simulations vs. database, respectively), for *Gender* (Male: 38% vs. 36%, Female: 62% vs. 64%), and for *Age* (Grade 2: 50% vs. 51%, Grade 3: 50% vs. 49%). The standard deviations around these means (expressed as a percentage of the mean) were also very small: for Educational Track (GE: 0.96%, TE: 0.73%, PE: 1.38%), for Gender (Male: 0.87%, Female: 0.53%), and for Grade (Grade 2: 0.72%, Grade3: 0.71%).

The next paragraphs report the condition means and the percentage of times that each effect was significant in the 1,000 samples, together with the minimum and maximum *z*-values (*t*-values for the effect of *Post length*).

3. Results

3.1. Collinearity between independent variables

We calculated the Variance Inflation Factor (VIF) in order to check whether there was no collinearity among the predictors. A GVIF test, the generalized version of VIF introduced by Fox and Monette (1992), was conducted based on a generalized linear model in which the binary response variable (correct/incorrect spelling of the verb ending) was predicted by all design factors: the social factors (*Gender*, *Age*, *Educational track*) and the (psycho)linguistic factors (*Paradigmatic <d> support*, *Bigram <d> support,*[17] *Whole-word frequency*, *PP family frequency*, *Post length*). The GVIF test indicated that multicollinearity was no cause for concern, as all squared generalized VIF scores were smaller than 2.24.[18] The conventional threshold for potentially problematic collinearity between predictors is 4 (Hair et al. 2010).

Despite these low VIF scores, we ran a linear model in which (the logarithm of) *Post Length* was predicted by the social variables, to avoid that a possible significant effect of these variables in the analyses on error probability (see below) could be due to *Post length* (see 1.2). There was a strong linear relationship between the length of the posts and *Gender* (ß = 0.11, SE(ß) = 0.02, $t = 6.28$, $p < .0001$) and *Age* (ß = 0.03, SE(ß) = 0.01, $t = 2.34$, $p < .02$).[19] Girls produced significantly longer posts than boys (t = [4.32 – 5.88], $p < .05$: 100.00%).[20] The average length of boys' posts was 14.38 tokens

[17] Recall that *Bigram <d> support* was operationalized in two ways: as bigram <d> support within the verb category only, and as bigram <d> support within all lexical categories other than verbs. Both bigram <d> support variables were used in the model-building phase.

[18] The GVIF values were normalized, taking the degrees of freedom into account, with the formula GVIF^(1/(2*df) – see Fox & Monette (1992).

[19] The same model building procedure was used as described in section 3.2. below.

[20] The means, the minimum and maximum t-values and the percentage of significance are based on 1,000 random samples.

vs. 17.17 tokens for girls' posts. Messages written by adolescents in Grade 3 (17.66 tokens on average) were significantly longer than the ones written by adolescents in Grade 2 (14.54) (t = [1.22 – 3.56], p < .05: 85.50%). This relationship between *Post length* and the social factors *Gender* and *Age* was also established by Hilte (2019: 159), which comes as no surprise since the posts that constitute the corpus for the present study were extracted from the corpus of Hilte (2019, see also Hilte et al. (2020a) for the social correlates of post length).

Although the GVIF test indicated that collinearity was no cause for concern, the strong linear relationship between *Post Length* and the social variables *Gender* and *Age* suggests that is advisable to use the residuals of the above linear regression model when building the statistical model.

3.2. Model building procedure

In order to analyze the binary response variable (correct/incorrect), we performed a generalized linear mixed-effects model, using the glmer function from the lme4 package (Bates, Maechler, Bolker, & Walker 2015) in the statistical software package R (R Core Team 2014). We started with a model in which *Chatter* was the only (random) variable and followed a stepwise forward procedure, each time adding one extra factor to the model.[21] First, we introduced *Lemma* as the second random factor, followed by each of

[21] There is no consensus about the best model-building procedure. Both stepwise forward, stepwise backward, and automated procedures have been advanced. Baayen (2014) advises a hypothesis-driven approach, which is the one taken here. Moreover, Hastie, Tibshirani, and Tibshirani (2017) conclude from their simulation data that "forward stepwise selection and best subset selection perform similarly throughout" (Hastie et al. 2017: 17).

the social variables and their pairwise interactions. Next, we added *Post length* and the psycholinguistic factors *Whole-word frequency, PP family frequency, Paradigmatic <d> support*, and *Bigram <d> support,*[22] and their interactions. Finally, the interactions between the social and the (psycho)linguistic variables were added one by one. In addition, the residuals of the linear model in 3.1, in which *Post length* was predicted by *Gender* and *Age*, was used as the predictor for *Post length*. This yielded the same results as when using the raw values of the predictors.

We used a likelihood ratio test ($\alpha = .05$) to decide whether adding an extra factor to the model accounted for significantly more variance in the data. If the complex model accounted for significantly more variance than the model without the extra factor in at least 75% of 100 simulations, the complex model was opted for. This procedure led to a final model in which the dependent variable (correct/incorrect spelling) was predicted by *Gender*, *Educational track*, *Age*, *Paradigmatic <d> support,* and *Bigram <d> support* within the verb category (as well as the two random factors). The variables *Post length*, *Whole-word frequency*, *PP family frequency,* and *Bigram <d> support* within lexical categories other than verbs were not included in the final model as they did not significantly improve the model. Henceforth, we will use the term *Bigram <d> support* to refer to the <d> support in *verb*-final bigrams (unless otherwise specified). The results presented in Table 3 and reported below are based on 1,000 simulations of this final model.

(INSERT TABLE 3 ABOUT HERE)

[22] Bigram <d> support within the verb category and bigram <d> support within lexical categories other than verbs were added as distinct variables to the model.

3.3. Overall error risk

From the perspective of a previous study on the spelling of verb homophones in adolescent social media writing (Surkyn et al. 2020), we note that the overall error rate for partially homophonous past participles appears to be relatively low: Only 11% of the past participles were spelled incorrectly, i.e., with a <t> ending instead of a <d> ending, while the error rate for the spelling of past participles with identical homophones was 29% (24.53% for past participles whose correct spelling was <d>, as in the present study). This major discrepancy can be explained by the absence of identical homophones for the past participles in this study, since precisely these identical homophones tend to cause a lot of confusion and trigger mistakes when the incorrect homophonous variant pops up (see e.g., Sandra et al. 1999). Moreover, the production of the final <d> for the past participle might be supported by the strong predictive relationship between the prefix *ge-*, which marks the past participle, and the suffix *-d*: Of all past participles in CELEX with the prefix *ge-* and the suffix sound [t] in final position (Baayen, Piepenbrock, & Gulikers 1995), 71.45% (type count, n = 871 out of 1,219) or 59.09% (token count, n = 4.098 out of 6,935) are spelled with <d>. This means that 7 out of 10 verbs and 6 out of 10 tokens trigger a <d> spelling for the past participle. Consequently, the past participle prefix *ge-* is significantly more strongly associated with the <d> spelling than with the <t> spelling ($X^2 = 229.29$, $p < .001$). This, too, might explain why the error rate is relatively low.[23]

[23] Note that autocorrection on Facebook Messenger and WhatsApp could also have contributed to the relatively low error rate on this type of past participles. However, this impact should definitely not be overestimated (see the General Discussion for more information).

3.4. Model output

The social variables *Gender, Educational track* and *Age* significantly affected the error rates (and did not interact). Girls made significantly fewer errors than boys (z = [-3.28 – -1.62], $p < .05$: 96.7%).[24] The former misspelled 9.90% of all past participles, on average, the latter 13.76%. Students in the more technical (TE) and practice-oriented (PE) tracks of secondary education made significantly more errors than their peers in the more general theoretical track (GE) (z = [3.97 – 5.22], $p < .001$: 100% and z = [4.72 – 5.99], $p < .001$: 100%, respectively). The errors rates for the TE and PE groups were highly comparable (z = [0.49 – 2.14], $p > .10$: 88.7%). The corresponding error percentages were: 3.57% (GE), 15.37% (TE), and 14.25% (PE). Furthermore, the older students made significantly fewer errors than the younger ones (z = [-3.61 – -1.81], $p <$ .05: 96.8%). The error rate in Grade 2 was 12.44% vs. 10.30% in Grade 3.

In addition to an effect of the social variables, we found a significant effect of the psycholinguistic factors *Paradigmatic <d> support* and *Bigram <d> support*. Significantly fewer errors were made as the paradigmatic <d> support increased (z = [-3.18 – -2.67], $p < .01$: 100%). Similarly, the higher the bigram <d> support, the fewer errors (z = [-5.50 – -5.00], $p < .001$: 100%). In other words, the larger the relative proportion of the <d> forms as opposed to the <t> forms in the inflectional paradigm and in the verb-final bigrams, the smaller the number of errors the adolescents made. Finally, the social factors and the frequency variables did not interact.

[24] Recall that the z-values reported here are the minimum and maximum z-values in the 1,000 samples, followed by the percentage of times that the effect was significant in the 1,000 samples.

Note that these effects were not driven by the past participles that occurred only once in the dataset. One might expect that (all or some) social variables only manifest themselves when very low-frequency forms are spelled (forms that occur only once in the corpus are a proxy for this), but the results remained the same, whether we analyzed all past participles or left these low-frequency ones out of the analyses.

As *Bigram <d> support* is a low-level, i.e., non-linguistic, factor whose effect was not demonstrated in earlier research on this type of spelling errors, we wanted to determine whether this new factor was sufficiently robust. More particularly, we checked whether it might be due to the high corpus frequency of a small number of items, i.e., the past participles whose corpus frequency exceeded 100 tokens. For these items, the value of this bigram measure recurred more than hundred times in the data set, which may have inflated its effect. Even though we addressed the problem of overrepresentation of these items by using a resampling technique, we performed an additional analysis in which we removed these items altogether. For the items with high corpus frequencies, this should dampen the effect of repeating the same value for each frequency variable. Even though this reduced our data set by more than half (i.e., 2,293 remaining observations), the effect of *Bigram <d> support* remained highly significant (z = -4.83, p < .001) as did the effect of *Paradigmatic <d> support* (z = -2.76, p = .006). As in the main analysis, neither the frequency of the past participle nor the frequency of the PP family was significant. We consider this as strong support in favor of two separate sources of <d> support: one based on bigram frequencies and the other on paradigmatic analogy.

Finally, we addressed the overrepresentation of some target forms by running a linear model in which the items' log odds were used as the dependent variable, i.e., the

logarithm of the proportion incorrect spellings of a past participle divided by the proportion correct spellings of this past participle (i.e., logits). The past participle *gezegd* and its compound past participles were included in this analysis, as a simple linear model based on logits at the item level should tackle the problem of the high occurrence frequency of a couple of verb forms as *gezegd*. Since some target forms were never misspelled, these logits could not be calculated, as $\log_{10}(0/1)$ yields infinity as a result. The value -2 was assigned to these items because the logarithm of the closest extreme odds (which never occurred in our dataset), i.e., .01/.99, yields a logit of -1.996. We did the same for items that were always spelled incorrectly, which yielded division by zero in the logit calculation $\log_{10}(1/0)$. The predictors in the final model were the same as in the mixed model: The only significant predictors were *Paradigmatic <d> support* (t = 1.90, p = .058) and *Bigram <d> support* (t = 3.84, p = .0001). The p-value of the former effect comes so close to .05 that it should be taken seriously, especially in the light of our entire set of results; *Paradigmatic <d> support* is highly significant in the mixed model and in the separate analyses of target forms with corpus frequencies above 1 and below 100 (see above).[25]

To find out whether the bigram variable reflected the type or token frequencies of the bigrams ending in <d> and <t> (the two variables were highly correlated, around .60 depending on the measure, see above), we ran the same model but used the logits of the bigram's type frequencies as a predictor. This type-based predictor fell far short of significance (p > .50). To make sure that the token measure of *Bigram <d> support*

[25] Especially in this context, it would be a mistake to treat .05 as a magical cut-off, as this was never even intended by Ronald Fischer, who is generally considered the 'father of modern statistical inference' (see Cumming 2012, for an insightful in-depth discussion on the use of p-values).

remained significant when it was decorrelated from type-based *Bigram <d> support*, we predicted the logits of the verb forms accuracy scores with (a) the residuals of a linear model in which (token-frequency based) *Bigram <d> support* was predicted by type-frequency based *Bigram <d> support* and (b) *Paradigmatic <d> support.* Even the residuals of the token-based measure remained highly significant (p = .0001). As in the model where the token measure of *Bigram <d> support* was not decorrelated from its type counterpart, *Paradigmatic <d> support* remained on the verge of significance (p = .066).

Thus, we controlled for the overrepresentation of some target forms in multiple ways: by (a) using a resampling technique, (b) an analysis without items that occurred only once, (c) an analysis without items that occurred very often, and (d) a simple linear model that predicts the log odds of the target forms' spelling correctness. In all these analyses we reached the same conclusion regarding the importance of two variables involving <d> support: one based on the frequency relationship between the inflectional variants within the verbs' inflectional paradigm (*Paradigmatic <d> support*) and one based on the frequency relationship between the final bigrams in inflected verb forms (*Bigram <d> support*).

We also performed additional analyses for the social variables. In addition to the mixed effects analyses we tested how the social factors affected the probability that adolescents made no errors across all target forms they produced. Even though we found an overall error rate of 11% on the target forms, quite a few chatters made no errors at all. As a matter of fact, 464 of the 629 chatters that remained after the *gezegd* forms were removed made no errors (73.77%). While only 1 out of 4 chatters spelled some or all their PPs incorrectly, they contributed more than half of the forms in the

database (2,696 out of the 4,978, 54.16%). We wanted to find out whether the social factors affected the probability that a chatter made no errors. As this probability was also related to the chatter's total number of tokens, we included the total number of targets per chatter as a covariate in the analysis (*TotalPPs*).[26] This number was indeed significantly related to *Educational track*: TE chatters contributed significantly more PPs than GE and PE chatters (mean: 14.13, 8.61 and 9.55, respectively; reference level TE, GE: ß = -5.88, SE(ß) = 1.84, t = -3.20, p = .002; PE: ß = -5.00, SE(ß) = 2.06, t = -2.43, p < .02), who did not differ from each other (t < 1). *Gender* had a significant effect, too: the mean number of PPs that girls contributed was higher than for boys (12.30 vs. 9.14; ß = 3.70, SE(ß) = 1.61, t = 2.29, p < 0.03). *Age* did not affect the number of PPs that a chatter produced (no improvement of the model in a ratio likelihood test: p = 0.61): Grade 2 and Grade 3 chatters produced about the same number of target forms (11.07 and 10.85, respectively).

The logistic model in which the probability that a chatter's PPs were all correct showed a significant effect of *Educational track* and *TotalPPs*. GE chatters were significantly more likely to spell all PPs correctly than the chatters in TE and PE (reference level GE, TE: ß = -1.18, SE(ß) = 0.28, t = -4.27, p < .0001, PE: ß = -1.44, SE(ß) = 0.30, t = -4.83, p < .0001), who did not differ from each other (t < *1*). Not surprisingly, the probability of making no errors decreased as the total number of a chatter's PPs increased (ß = -0.05, SE(ß) = 0.01, t = -5.23, p < .0001). *Gender* and *Age*

[26] We excluded chatters who contributed only one past participle to see whether those who spelled all forms correctly were able to spell these forms correctly more than once. Although this did not affect the results, we report the data on the chatters who provided at least two forms.

had no effect, in contrast to what we found in the mixed model analyses. The importance of these results will be discussed in the General Discussion.

4. General Discussion

In this study, we examined unintentional spelling errors on partially homophonous past participles in informal online chat conversations of Flemish teenagers. Previous psycholinguistic research on Dutch verb spelling errors primarily investigated *identical* homophones (e.g., Bosman 2005; Sandra et al. 1999; Schmitz et al. 2018; Surkyn et al. 2020). In the present study, however, we shifted our focus to past participles that are only partially homophonous with other forms in the verbs' inflectional paradigm.

Our results showed that the spelling production of the past participles was affected by the token distribution of <d> and <t> spellings in the suffixes of analogical sets of inflected forms, i.e., partially homophonous and non-homophonous (morphological) variants (e.g., *noemt,* 'calls', *noemde,* 'called', for *genoemd,* '(has) called'). There was a significant effect of the frequency rates of these relevant forms: A greater <d> support in the verb's inflectional paradigm caused fewer errors. Conversely, higher <t> support in the form of higher frequencies of the homophonous present tense forms ending in <t> led to a higher error rate. This result makes sense, as the correct spelling of the past participles in this study requires a <d> ending. Hence, the frequency of morphologically related forms (i.e., other forms in the verb's inflectional paradigm such as the simple present and simple past) has an analogical effect on the spelling of partially homophonous past participles. Thus, the intraparadigmatic analogical effects attested by Ernestus and Mak (2005) in speeded word reading experiments with Dutch verb forms are now also established in a natural writing context.

Moreover, the analyses revealed an effect of *Bigram <d> support* within the verb category: The more frequent a final bigram ending in <d> (e.g., <rd>) occurred in inflected verb forms compared to the verb-final bigram ending in <t> (e.g., <rt>)[27] (i.e., the higher the value of this ratio), the fewer errors were made on the partially homophonous past participles spelled with a <d> ending. This finding, i.e., an effect of *Bigram <d> support*, is consistent with the findings of Gahl & Plag (2019). In their study, they observed an effect of the variant bigram (i.e., the incorrectly spelled bigram) on the spelling of English derivational suffixes produced in tweets: The higher the probability of the variant bigram, the higher the probability that the spelling variant (i.e., a spelling error) occurred in the Twitter corpus (Gahl & Plag 2019: 20). In the present study, a greater <t> support (i.e., a smaller <d> support) in the verb-final bigrams also caused more errors. Hence, besides an analogical effect at the paradigmatic level, our study reveals an analogical effect at the sublexical level. An effect of homophony at the sublexical level was already observed in earlier research by Sandra (2010) and Sandra & Van Abbenyen (2009) (see Section 1.1 Psycholinguistic variables). However, in these studies, the homophonous letter strings always involved the same suffix spelling and one vs. two stem-final letters (e.g., *waste,* 'washed': *s* + *te* vs. *tastte,* 'groped': *st* + *-te*). In contrast, our measure of *Bigram <d> support* involves the two final letters of a verb form, which do not necessarily include a suffix letter. For instance, even though both *antwoord,* 'answer', and *gebeurd,* 'happened' contribute to the frequency of <rd>, the <d> is the stem-final letter in the former example but the suffix letter in the latter.

[27] Recall that the bigram pairs (e.g., <rd> and <rt>) are homophonous: Both the final <t> and <d> are pronounced as [t] (due to final devoicing).

Hence, *Bigram <d> support* is a factor at an even lower level than the sublexical factors reported earlier.

The frequency of the word-final bigrams in words of lexical categories other than verbs had no significant impact on the spelling of the past participles. This result indicates that the final bigrams in nouns as *moord* ('murder'), *woord* ('word') and *soort* ('kind') and in adjectives as *hard* ('hard'), *absurd* ('absurd') and *kort* ('small') do not affect the spelling of past participles as *geleerd* ('learned') and *gehoord* ('heard'). Hence, the *Bigram <d> support* factor is not a purely orthographic factor, for then its effect would not be limited to the lexical category of verbs. Instead, it reflects morphologically conditioned orthographic patterns. Recall that the spelling of the past participle is modeled on the spelling of the past tense: A stem-final voiced consonant selects the past tense allomorph *-de* (spelled as <de>) and, hence, the past participle allomorph *-d* (spelled as <d>). Thus, within the set of verbs, voiced stem-final consonants give rise to a <d> spelling of the past participle suffix (in all verbs), hence increasing the <d> support for (partially homophonous) past participles whose final bigram begins with one of these letters. The consequence of this analogical modeling of the past participle on the past tense is that a final bigram like <rd>, which is the correct spelling in these past participles, looks more acceptable. The analysis revealed that the bigrams' token frequencies and not their type frequencies are the determinant of the error risk. Thus, *Bigram <d> support* is an orthographic factor emerging from the morphographic spelling of Dutch.

The finding that both *Paradigmatic <d> support* and *Bigram <d> support* affect the error risk on the spelling of (partially homophonous) past participles emphasizes the fact that morphological relations affect the spelling of the past

participles in two different, yet independent ways. Note that the correlation between these two factors was very small: .21, i.e., their shared variance is only 4.4%.

The effect of *Whole-word frequency*, i.e., the frequency of the target form itself, and *PP family frequency*, i.e., the summed token frequencies of the target past participle and all forms containing the base past participle, were not significant. The fact that these two frequency variables did not affect the spelling of the past participles in our study, but the <d> support variables did, indicates that morphological relations are the major driving forces behind the spelling of these partially homophonous past participles, whose suffix is spelled as <d>. These are (a) inflected forms of the same verb and (b) the morphographically conditioned final bigrams that end in <d> or <t> of the inflected forms across all verbs. This means that partial and sublexical homophones ending in <d> *and* <t> (within the lexical category of verbs) play an important role. For example, the spelling of the past participle *geleerd* ('learned') is not only affected by (the frequency of) *geleerd* (the past participle itself) and *aangeleerd* ('taught') and *afgeleerd* ('unlearned') (the compound past participles containing the base PP), but also by the partial homophone *leert* ('learns', simple present) and by *leerde* ('learned', simple past), as well as by the frequency of the final bigram of other inflected verbs as *word* ('become'), *vermoord* ('kill'), *hoort* ('hears') and *probeert* ('tries'), which are homophonous with the final bigram of *geleerd* at the sublexical level.

Furthermore, the unique database made it possible to include the impact of the social factors *Gender*, *Educational track* and *Age* in the research design. Our results revealed a significant effect of these three social factors on the error rates of partially homophonous past participles. Girls made fewer errors than boys, older teenagers (Grade 3) outperformed the younger ones (Grade 2) and adolescents in general

education (GE) showed fewer errors than their peers in technical and professional education (TE and PE), who did not differ from each other. These results are consistent with our hypotheses and correspond to previous research on the spelling of identical verb homophones in adolescent online writing (Surkyn et al. 2019, Surkyn et al. 2020).

Whereas the above findings are made from the perspective of single targets, the role of the social factors was also assessed in analyses where individual chatters were the unit of analysis. Whereas the former analysis focuses on the probability that a random target form in our database is spelled incorrectly, the latter measures the probability that none of the target forms produced by a chatter are spelled incorrectly. In these analyses, only *Educational track* was significant: GE chatters were more likely to make no errors than TE and PE chatters, when the effect of the total number of PPs was entered as a covariate in the analysis. Moreover, the total number of PPs in the set of error-free chatters did not differ between the three educational tracks.

Together, the results from these two types of analyses indicate that the effects of *Gender* and *Age* should be interpreted with caution. On the one hand, it is clear that not all boys produce more errors (on these target forms) than girls and that not all younger adolescents make more errors than older ones. On the contrary, the probability that a chatter makes no errors does not differ between boys and girls, nor does it differ between chatters in Grades 2 and 3. On the other hand, the significant effects of *Gender* and *Age* in the mixed model analyses indicates that, if boys or younger teenagers make one or more errors on these past participles, they make more errors on average than their female peers or older teenagers, respectively. In other words, the effects in the mixed models are carried by subsets of boys and younger chatters. The finding that *Educational track* had a significant effect in both types of analyses indicates that this

factor is a very powerful one. Not only is it more probable that an error occurs in the chats of TE and PE adolescents (the latter do not differ) than in the chats of their GE peers, it is also more probable to find a chatter without any errors in the GE group than in the TE and PE groups (the latter do not differ).

The finding that students in general education outperformed their peers from technical and professional education is in line with our hypothesis. Adolescents in the more theory-oriented tracks are likely to have a better knowledge of verb spelling rules and better metalinguistic skills, since their official curricula focus more on (reflection on) spelling issues than the curricula for more practice-oriented tracks (VVKSO 2006, 2014). This good rule knowledge may go hand in hand with a more outspoken attitude to avoid these errors, such that the effect of *Educational track* manifests itself as (a) a smaller probability that an error is made on a randomly selected PP when the chatter belongs to the GE group than to the TE or PE groups and (b) a larger probability that a chatter who makes no errors on the set of PPs that she or he contributes to the corpus is a GE chatter than a TE or PE chatter. It seems likely that differences in both spelling proficiency and attitudes towards avoiding verb spelling errors account for our finding that *Educational track* is the strongest social error determinant.

Besides differences in the educational trajectories of GE, TE, and PE chatters, there might be differences in their exposure to correct and incorrect spellings of the past participles (and the partially homophonous forms in their inflectional paradigm). Students in the more theory-oriented tracks are likely to have more experience with printed words (e.g., newspapers and novels) and proof-read texts, and hence with (formal) standard writing. This, in turn, may have a positive impact on the quality of their orthographic representations (The Lexical Quality Hypothesis, see e.g., Perfetti &

Hart 2002, and Perfetti 2007). In contrast, students in more practice-oriented tracks might be subject to a proportionately higher exposure to 'misspelled' word forms because their exposure to (formal) standard writing might be minimal and consequently their relative exposure to social media writing might be higher. This relation between individuals' experience with printed texts on the one hand and the quality of orthographic representation of these individuals on the other hand has been studied, for example, in experimental research on orthographic learning and spelling errors in particular (see e.g., Falkauskas & Kuperman 2015; Kuperman et al. 2021; Ouellette, Martin-Chang, & Rossi 2017; Rossi, Martin-Chang, & Ouellette 2019). Hence, the quality of our chatters' orthographic representations may be strongly correlated with *Educational track.*

An analogy can be made here with a completely different factor: the home language of the youngsters. Among the teenagers from technical and professional education, Dutch appears to be much more often not (one of) the home language(s) than among the teenagers from general education (Hilte et al. 2018: 9). However, we are not able to capture the impact of home language and exposure to formal writing (separately) in our study. Previous analyses on the social profile of the adolescents in our corpus showed that educational track and socio-economic background are highly correlated (see Hilte et al. 2018). In other words, home language is not only related to educational track, but also to the socio-economic background of the youngsters (determined on the basis of the profession of their parents). And the same might hold for the exposure to formal writing. So, these factors are completely intertwined and therefore hard to disentangle. However, this does not invalidate the conclusion that both minimal

exposure to formal standard writing and minimal exposure to the Dutch language in home contexts most probably reinforced the effect of *Educational track*.

Crucially, this effect of *Educational track* did not interact with the psycholinguistic variables *Paradigmatic <d> support* and *Bigram <d> support* (nor did the other social factors). This suggests that, despite plausible differences in the quality of the orthographic representations between the three educational groups, the frequency distributions of correct and incorrect spellings of the target forms and the partial homophones in their inflectional paradigm were sufficiently comparable across the different groups.

The effect of *Gender* can be related to the demonstration in many sociolinguistic studies that women tend to show a stronger norm sensitivity and that they try to avoid stigmatized forms more than men (see e.g., Labov 2001: 293; Tagliamonte 2011: 32). Stigmatization is an issue here, since the public discourse related to these errors tends to be very stigmatizing (Verhaert & Sandra 2016). While one could argue that youngsters might be more careless about verb spelling in private informal online writing, female adolescents still appear to show more error-avoiding behavior than their male peers. Recall that this effect of *Gender* was observed in an analysis at the level of individual 'responses', i.e., when individual target forms were treated as the unit of analysis, whereas the effect was absent when the (binary) dependent variable was the faultless spelling of all target forms produced by the same chatter, i.e., when chatters were the unit of analysis. Given this difference in the unit of analysis, there is no contradiction here. The former analysis indicates that, on average, a randomly selected error on these past participles is more likely to be have been made by a boy than by a girl. The latter analysis indicates that there are as many boys as girls who do not make these errors at

all. Hence, the conclusion that boys make more errors than girls does not mean that *all* boys produce more errors than *all* girls. On average, they indeed do – and this is what the fine-grained analysis reveals – but generalizing this finding to all members of the group would be an error of inference. We conclude that girls, on average, adopt a (spelling) attitude that reflects more rule-obedience and, hence, error-avoidance than boys, which is in keeping with our earlier findings (Surkyn et al. 2019; Surkyn et al. 2020).

In Surkyn et al. (2019) we suggested that the observed effect of gender could also be an effect of post length in disguise. We hypothesized that boys might have a stronger focus on speed, since they were found to produce shorter messages than girls (see also Hilte 2019: 159; Hilte et al. 2020a). When time pressure causes extra pressure on working memory resources needed for correct rule application, the error risk increases (Sandra et al. 1999). The corpus for the present study, which is extracted from the Hilte-corpus, unsurprisingly also revealed an effect of *Gender* on the average post length, and so it did for *Age* (see also Hilte et al. 2020a): Girls produced longer posts than boys and the older adolescents wrote longer messages than the younger ones. However, our analyses showed that *Post length*, after being statistically decorrelated from these social factors, had no impact on the error rates. This is highly relevant for the interpretation of the effect of *Gender*, since it means that it is not a secondary effect of *Post length*.

Finally, for the effect of *Age*, i.e., younger adolescents (Grade 2) make more errors than older ones (Grade 3), we refer to sociological and sociolinguistic findings with respect to teenage behavior: It has been observed that non-conformist behavior peaks amongst teenagers in their mid-teens and decreases in their late teens, as they

approach adulthood (Holmes 1992: 184; Coates 1993: 94; Tagliamonte 2016). In other words, younger teenagers can be expected to be more rebellious with respect to standard writing norms. While deviant verb spellings in Dutch are no obvious tools for gaining peer group prestige (as many spellers and readers are not aware of the errors), younger adolescents might attach less importance to spelling conventions in general and consequently also invest less energy in the application of the verb spelling rules. Hence, like the gender effect, this effect might be due to a more negligent attitude of (young) spellers, and, to a lesser extent than for *Educational track*, to poorer rule knowledge (although a recent study of Chamalaun and colleagues (in press) showed better grammatical knowledge in older adolescents). Note that the effect of *Age* (like the effect of *Gender*) was unaffected by *Post length*. Apart from that, the older students in Grade 3 most probably have more experience with printed words and proof-read texts, and thus with (formal) standard writing than the younger students. Hence, the same remarks concerning potential differences in the quality of the orthographic representations in the different groups (see above for the discussion of the effect of *Educational Track*) can be made here with respect to the established age differences. However, the same remark applies as above: Despite possible differences at the representational level, we found no interaction between this social variable and the psycholinguistic variables involving <d> support. Here, too, this suggests that the relevant part of the orthographic representational system of the two age groups was sufficiently similar for our purpose.

Furthermore, our results revealed no interactions between the social variables and the psycholinguistic frequency variables. Thus, all social groups seem to benefit equally from the effect of *Paradigmatic <d> support* and *Bigram <d> support*, irrespective of their error rates. This finding supports our hypothesis (which was based

on the model of Sandra and colleagues 1999), i.e., as differences between the groups on a social factor will be due to differences in rule knowledge and/or the willingness to observe these spelling rules (attitude), such a factor will affect the conscious process that determines the suffix spelling in working memory. Hence, it is expected to have an effect on the error rates but not on the nature of the errors, which reflects automatic processes underpinning the retrieval of orthographic forms. The latter are driven by frequency, which governs a form's accessibility, and analogy, which determines the coactivation of orthographic representations, thus yielding an analogical set. The absence of such interactions is also consistent with the results from our previous study on verb spelling errors on homophonous verb forms and the interplay between social and psycholinguistic factors (Surkyn et al. 2020).

As mentioned in section 2.3., the past participle *gezegd* ('said') and its compounds were excluded from the analyses because of its atypical behavior: while *gezegd* is far more frequent than any other past participle, both in our corpus (1,573 of all 6,551 target forms, i.e., 24%) and in the SUBTLEX-NL database (i.e., with 404 occurrences per million words it is the second most frequent past participle, after *gebeurd* ('happened'), with 438 tokens per million),[28] it has an exceptionally high error rate. This is quite surprising, since we might expect that this high frequency leads to a very strong representation of the correct form. However, another frequency effect might be at work here. As the incorrect form *gezegt* (instead of *gezegd*) occurs exceptionally often in online chat conversations (both as a base form and in compound verbs) (26.57%), implicit learning of the incorrect spelling might lead to the storage of this alternative orthographic representations (see e.g., Rahmanian & Kuperman 2019) or,

[28] The five compound forms of *gezegd* are all highly infrequent (range: 0.02 – 4.28 per million).

alternatively, to purposeful imitation in certain groups of chatters. In other words, our data might suggest that this particular misspelling has become a typical marker of the online writing of some chatters, though it is unclear to what extent the incorrect spelling is applied deliberately. The fact that the effects of the social factors on *gezegd* are not the same as for the other past participles (in particular, the absence of a gender effect, see below) might indicate that a different mechanism is at work for this form and its compounds. Additional analyses showed that the probability of making an error (i.e., *gezegt*) is predicted by *Educational track*, the error risk being significantly higher for TE chatters than for GE chatters (reference level GE, TE: ß = 3.88, SE(ß) = 1.43, t = 2.72, p = .007) whereas the error risk of PE chatters did not differ from either group (reference level GE, PE: ß = 2.07, SE(ß) = 1.51, t = 1.37, p > .10; reference level TE: PE: ß = -1.81, SE(ß) = 1.25, t = -1.44, p > .10). The error risk was highest for Grade 2 chatters (ß = -3.37, SE(ß) = 0.55, t = -6.08, p < .0001). The effect of *Gender* did not improve the logistic regression model (likelihood ratio test: p = .34). The same two factors were also significant when measuring the probability that a chatter spelled all forms of *gezegd* correctly, when the total number of these forms in chatters' contributions was used as a covariate. The effect of *Educational track* showed that this probability was significantly larger for GE chatters than for their peers in both TE and PE (TE: ß = -2.86, SE(ß) = 0.63, t = -4.52, p < .0001; PE: ß = -2.72, SE(ß) = 0.66, t = -4.14, p < .0001), who did not differ from each other (t < 1). Students in Grade 2 were less likely to spell all these past participles correctly than students in Grade 3 (ß = 0.81, SE(ß) = 0.34, t = 2.41, p < .02). Again, *Gender* did not significantly improve the model (p = .89). Even though we can only speculate about the interpretation of these results, the combination of a high error rate for the *gezegd* forms and a different impact of the

social factors might be indicative of a tendency to imitate errors on this high-frequency form, which occurred more in some social groups than in others. If so, it is unclear whether this case can be lumped together with the (abundantly present) deliberate spelling manipulations we encounter in the corpus: e.g., respelling of the cluster <ch> into <g> (*sgattig* instead of *schattig*, 'cute') (see also Hilte 2019: 180).

Furthermore, by studying *partially* homophonous past participles, we were able to compare our results with the results from our previous study on identical verb homophones (Surkyn et al. 2020). Adolescents made remarkably fewer errors on the partially homophonous past participles in this study (11%) than on the fully homophonous past participles in our previous study (29%, Surkyn et al. 2020, Study 2). This suggests that the impact of partially homophonous verb forms such as *noemt* ('names') on the spelling of *genoemd* ('named') is less strong than the impact of an identical homophones such as *wordt* ('becomes') on *word* ('become'). Even though intraparadigmatic analogy is at work (see the effect of *Paradigmatic <d> support* established in the present study), the coactivation that is triggered by these partially homophonous verb forms does not seem to cause the same kind of pop-ups of misspellings as identical verb homophones.

Finally, we should note that, in principle, autocorrection on Facebook Messenger and WhatsApp could also have contributed to the relatively low error rate on this type of past participles. However, we are confident that the impact of autocorrection on the data is limited. First of all, the corpus contains, in addition to these non-existent spellings of past participles, many other alternative spellings that tend to be corrected automatically if autocorrection is activated. Many of these spellings are prototypical markers of the genre (e.g., highly frequent but non-standard abbreviations: *gwn* instead

of *gewoon* ‘just’), others are ad hoc creative spelling manipulations or unintended spelling errors of all kinds. The fact that these non-standard spellings are abundant in the corpus (and produced by all social groups) (see Hilte et al. 2020b) indicates that most adolescents switch off or ignore autocorrection (regardless of their gender, age and educational track). Moreover, if the spelling of these past participles was strongly influenced by autocorrection, we would not be able to uncover such systematic social and psycholinguistic patterns.

5. References

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard languages and language standards in a changing Europe* (pp. 145–161). Oslo: Novus.

Baayen, R. H. (2014). Multivariate Statistics. In R. Podesva, & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 337-372). Cambridge: Cambridge University Press.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67*: 1-48.

Bosman, A. M. T. (2005). Development of rule-based verb spelling in Dutch students. *Written Language & Literacy 8*: 1-18.

Chamalaun, R., Bosman, A., & Ernestus, M. (in press). The Role of Grammar in Spelling Homophonous Regular Verbs. *Written Language & Literacy.*

Coates, J. (1993). *Woman, men and language. A sociolinguistic account of sex differences in language*. London: Longman.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge / Taylor & Francis Group. differences in language. London: Longman.

Ernestus, M., & Mak, W. M. (2005). Analogical effects in reading Dutch verb forms. *Memory & Cognition 33*: 1160-1173.

Falkauskas, K., & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition 41*: 1607.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *JASA*, *87*: 178–183.

Frisson, S., & Sandra, D. (2002). Homophonic Forms of Regularly Inflected Verbs Have Their Own Orthographic Representations: A Developmental Perspective on Spelling Errors. *Brain and Language 81*: 545-554.

Gahl, S., & Plag, I. (2019). Spelling errors in English derivational suffixes reflect morphological boundary strength: A case study. *The Mental Lexicon 14*: 1-36.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2010). *Multivariate Data Analysis. 7th ed.* New York: Pearson.

Hastie, T.J., Tibshirani, R., & Tibshirani, R.J. (2017). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso.

Hilte, L. (2019). *The social in social media writing: The impact of age, gender and social class indicators on adolescents' informal online writing practices*. Antwerp: University of Antwerp (doctoral thesis).

Hilte, L., Daelemans, W., & Vandekerckhove, R. (2020a). Lexical patterns in adolescents' online writing: The impact of age, gender, and education. *Written Communication 37*: 365-400.

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018). Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics 7*: 2-25.

Hilte, L., Vandekerckhove, R. & Daelemans, W. (2020b). Modeling adolescents' online practices: The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik 87*: 173-201.

Holmes, J. (1992). An introduction to sociolinguistics. London / New York: Longman.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods 42*: 643-650.

Kuperman, V., et al. (2021). Prevalence of spelling errors affect reading behavior across languages. *Journal of Experimental Psychology General*, 1-61.

Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Malden, MA: Blackwell Publishers.

Ouellette, G., Martin-Chang, S., & Rossi, M. (2017). Learning from our mistakes: Improvements in spelling lead to gains in reading speed. *Scientific Studies of Reading*, *21*: 350-357.

Perfetti, C. A., (2007). Reading ability: Lexical quality to comprehension. *Scientific studies of reading 11*: 357-383.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of functional literacy 11*: 67-86.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahmanian, S., & Kuperman, V. (2019). Spelling errors impede recognition of correctly spelled word forms. *Scientific Studies of Reading 23*: 24-36.

Rossi, M., Martin-Chang, S., & Ouellette, G. (2019). Exploring the space between good and poor spelling: Orthographic quality and reading speed. *Scientific Studies of Reading 23*: 192-201.

Sandra, D. (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*: 405-444.

Sandra, D., & Van Abbenyen, L. (2009). Frequency and analogical effects in the spelling of full-form and sublexical homophonous patterns by 12 year-old children. *The Mental Lexicon 4*: 239-275.

Sandra, D., Frisson, S., & Daems, F. (1999). Why simple verb forms can be so difficult to spell: the influence of homophone frequency and distance in Dutch. *Brain and Language 68*: 277-283.

Schmitz, T., Chamalaun, R., & Ernestus, M. (2018). The Dutch verb-spelling paradox in social media. A corpus study. *Linguistics in the Netherlands*, 111-124.

Schreuder, R, & Baayen, R.H. (1997). How Complex Simplex Words can be. *Journal of Memory and Language 37*: 118-139.

Surkyn, H., Vandekerckhove, R., & Sandra, D. (2019). Errors Outside the Lab: The Interaction of a Psycholinguistic and a Sociolinguistic Variable in the Production of Verb Spelling Errors in Informal Computer-Mediated Communication. In J. Longhi & C. Marinica (Eds.), *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019)* (pp. 59-62). Paris.

Surkyn, H., Vandekerckhove, R., & Sandra, D. (2020). From experiment to real-life data: social factors determine the rate of spelling errors on rule-governed verb

homophones but not the size of the homophone dominance effect. *The Mental Lexicon 15*: 422-463.

Tagliamonte, S. (2011). *Variationist Sociolinguistics: Change, observation, interpretation*. Oxford: Wiley-Blackwell.

Tagliamonte, S. (2016). *Teen talk: The language of adolescents*. Cambridge: Cambridge University Press.

Van den Bergh, H. & van Es, A., & Spijker, S. (2011). Spelling op verschillendeniveaus: werkwoordspelling aan het einde van de basisschool en het einde van het voorgezet onderwijs. [Spelling at different levels: verb spelling at the end of primary school and at the end of secondary education]. *Levende Talen Tijdschrift 12*: 3-14.

Van der Horst, M., van den Bergh, H., & Evers-Vermeul, J. (2012). Kunnen leerlingen wat ze moeten kunnen? Onderzoek naar de doorlopende leerlijn op het gebied van werkwoordspelling. [Can students do what they need to be able to do? Research into the continuous learning line in the field of verb spelling]. *Levende Talen Tijdschrift 2*: 33-42.

Vandekerckhove, R., & Nobels, J. (2010). Code eclecticism: linguistic variation and code alternation in the chat language of Flemish teenagers. *Journal of sociolinguistics 14*: 657-677.

Verhaert, N., & Sandra, D. (2016). Homofoondominantie veroorzaakt dt-fouten tijdens het spellen en maakt er ons blind voor tijdens het lezen. [Homophone dominance causes dt-errors during spelling and makes us blind to them while reading]. *Levende Talen Tijdschrift 17*: 37-46.

Verhaert, N., Danckaert, E., & Sandra, D. (2016). The dual role of homophone dominance. Why homophone intrusions on regular verb forms so often go unnoticed. *The Mental Lexicon 11*: 1-25.

VVKSO. (2006). *Leerplan Nederlands secundair onderwijs bso*. [Curriculum Dutch secondary education bso]. Brussel: Vlaams Verbond van het Katholiek Secundair Onderwijs. Retrieved from: http://ond.vvkso-ict.com/leerplannen/doc/Nederlands-2006-020.pdf

VVKSO. (2014). *Leerplan Nederlands secundair onderwijs aso-kso-tso*. [Curriculum Dutch secondary education aso-kso-tso]. Brussel: Vlaams Verbond van het Katholiek Secundair Onderwijs. Retrieved from: http://ond.vvkso-ict.com/leerplannen/doc/Nederlands-2014-001.pdf

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley Press.

Acknowledgements

We thank the two anonymous reviewers for their helpful and constructive feedback on the previous version of this paper.

Funding

The research reported here was supported by grant G023118N from the Research Foundation – Flanders (FWO).

Appendix

Table 1. *Data distribution of Gender, Educational track and Age*

Variable	Variable levels	Participants	Posts	Tokens	Target forms
Gender	**Boys**	631 (48.99%)	140,249 (34.76%)	774,530 (32.82%)	2,305 (35.19%)
	Girls	657 (51.01%)	263,242 (65.24%)	1,585,587 (67.18%)	4,246 (64.81%)
Educational track	**General Education (GE)**	559 (43.40%)	115,900 (28.72%)	711,076 (30.13%)	1,944 (29.67%)
	Technical Education (TE)	370 (28.73%)	185,546 (45.99%)	1,090,580 (46.21%)	3,109 (47.46%)
	Professional Education (PE)	359 (27.87%)	102,045 (25.29%)	558,461 (23.66%)	1,498 (22.87%)
Age	**Younger adolescents (Grade 2: 15-16)**	978[29] (52.16%)	213,761 (52.98%)	1,189,661 (50.41%)	3,298 (50.34%)
	Older adolescents (Grade 3: 17-20)	897 (47.84%)	189,730 (47.02%)	1,170,456 (49.59%)	3,253 (49.66%)

[29] Note that the sum of the number of the younger and older participants does not correspond to the total number of participants. That is because the same participants can occur several times (at different ages) in the corpus if they submitted both recent messages and older ones.

Table 2. *Forms in the inflectional paradigm affecting <d> support and <t> support*

Inflected form	**Morphological structure**	**Example**
<d> form		
past participle	*ge-* + stem + *-d*	*geleerd*
simple past, singular	stem + *-de*	*leerde*
simple past, plural	stem + *-den*	*leerden*
attributively used past participle	*ge-* + stem + *-d*	*geleerd*
inflected attributively used past participle	*ge-* + stem + *-de*	*geleerde*
present participle	stem + *-end*	*lerend*
inflected present participle	stem + *-ende*	*lerende*
<t> form		
2nd person singular simple present	stem + *-t*	*leert*
3rd person singular simple present	stem + *-t*	*leert*

Figure 1. *Frequency distribution of the past participles*

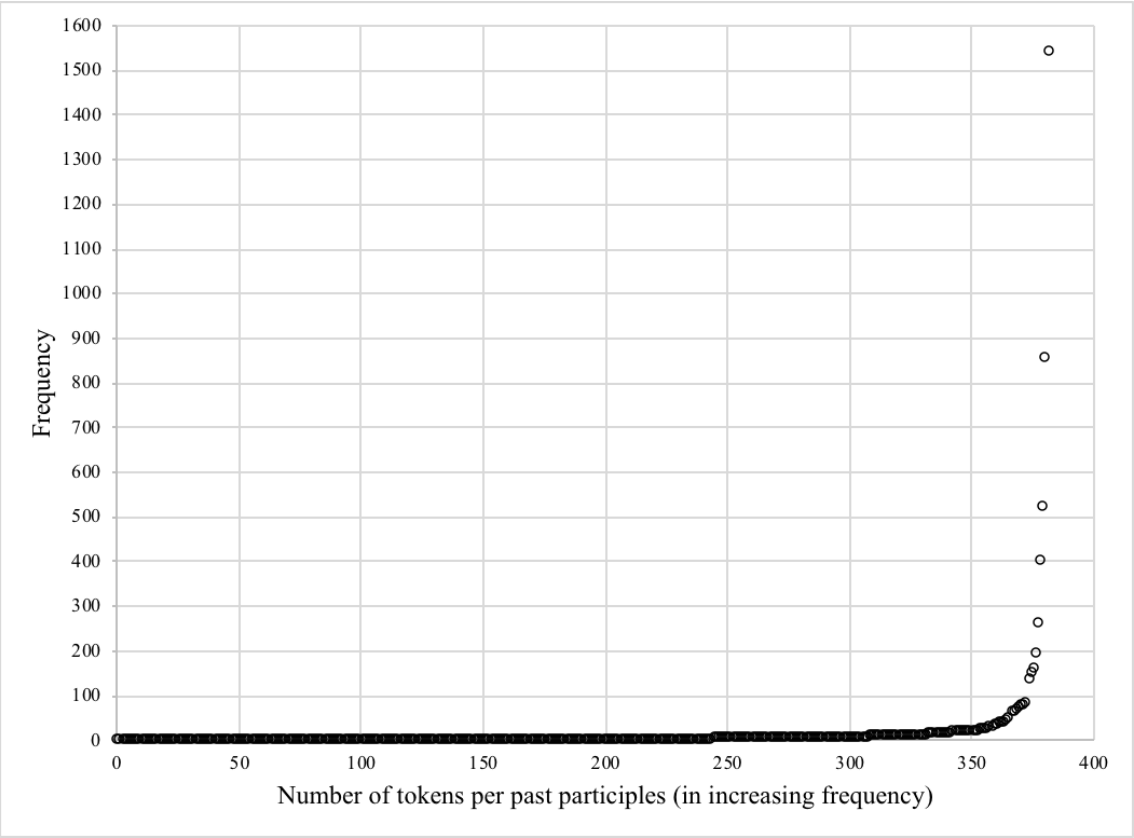

Table 3. *Minimum and maximum values for the regression slope (ß), the standard error (SE) of the slope, the associated z-value, and the number of times that the predictor was significant at different alpha levels in the 1,000 random samples*

Predictor	**ß**		**SE(ß)**		***z***		**Number of times below *p***		
	min.	**max.**	**min.**	**max.**	**min.**	**max.**	**.05**	**.01**	**.001**
Intercept: GE – Male – Grade2	-4.12	-3.53	0.31	0.38	-11.69	-10.44	-	-	1000
Paradig-matic <d> support	-1.00	-0.83	0.29	0.33	-3.18	-2.67	-	1000	0
Bigram <d> support	-1.47	-1.20	0.23	0.27	-5.50	-5.00	-	-	1000
Female (vs. Male)	-0.84	-0.38	0.23	0.27	-3.28	-1.62	582	385	0
TE (vs. GE)	1.22	1.57	0.28	0.34	3.97	5.22	-	-	1000
PE (vs. GE)	1.55	2.04	0.31	0.37	4.72	5.99	-	-	1000
PE (vs. TE)	0.14	0.62	0.26	0.32	0.49	2.14	14	0	0
Grade 3 (vs. Grade 2)	-0.69	-0.35	0.18	0.21	-3.61	-1.81	346	622	21

Note. Only the fixed predictors that were included in the final mixed model are presented. *Chatter* and *Lemma* were random intercepts in the model.