From the Valleys to the World Wide Web: Non-Standard Spellings on Social Network Sites

Aivars GLAZNIEKS
Eurac Research, Bolzano, Italy

Alexander GLÜCK
Free University of Bolzano, Italy

This paper focuses on non-standard spelling patterns in Facebook messages. We analyse German Facebook messages from around 100 South Tyrolean Facebook users collected for the DiDi Corpus (www.eurac.edu/didi) with respect to the variants used for unstressed word-final <-er> (pronounced as [ɐ]) in polysyllabic words in Standard German. The DiDi Corpus contains a considerable portion of non-standard and dialectal messages; in addition, the corpus provides a wide range of socio-demographic information about the writers (such as age, gender, geographic origin and other) which is available as metadata. The corpus thus is suitable to investigate spelling variation from a sociolinguistic point of view taking into account extra-linguistic features of the Facebook users. We will provide examples of regional features and analyse the distribution of such features. Results show that the standard is the most preferred variant in the corpus; however, age is a strong predictor for variation as young writers more often use an alternative spelling than older writers do. In addition, the non-standard variants in the corpus reflect features of spoken varieties and differences within the investigated territory. Thus, the geographic origin of writers is a good predictor for the way of variation.

1. Introduction

Social Network Sites (SNS) claim that they are “on a mission to connect the world”\(^3\). They facilitate communication among people wherever they are located. Consequently, many users of SNS communicate with a broad and heterogenic group of friends on different occasions and thereby express diversely various aspects of their identities (such as gender, age, ethnic background etc.). One aspect might also be their local identity, which users can show linguistically by the use of a regional variety. In doing so, the extent of dialect used in computer-mediated communication (CMC) may depend on various factors such as the individual dialect skills, the vividness and prestige of the respective dialect in the community, emotional involvement in the topic, age, gender, the intended audience, and other factors interacting with each other (Peersmann et al. 2015). Non-standard spellings in CMC are interesting for linguistic research for several reasons. Since no orthographic rules are usually available for non-standard varieties, it is up to the users to represent their dialect in a proper but readable and comprehensible way. Therefore, dialect spellings may not always be coherent (neither with respect to a group of dialect speakers nor with respect to individual

\(^3\) Mark Zuckerberg used this expression in a Facebook post published on 21\(^{st}\) of August 2013 in which he discussed connectivity as a human right (cf. https://www.facebook.com/zuck/posts/10100933624710391).
writers) but may appear in various appropriate forms (Siebenhaar 2005, Müller 2011). Thus, in representing their spoken variety in a written form, authors show their linguistic knowledge and awareness of non-standard features in a creative way, which are included in spontaneous writing. On the other hand, non-standard spellings reveal constraints of spontaneous writing, because the range of creativity is limited by existing orthographic rules, even if the writers do not completely adhere to these rules (cf. Sebba 2007: 26-57). What sounds like a paradox at first glance becomes comprehensible when you consider e.g. the sound-letter correspondences of non-standard spellings. In order to be readable and understandable by the intended audience, these spellings must refer to an existing system, which is in most cases the orthographic system of the respective standard language and which is shared by both the writer and the audience as common knowledge. However, since CMC exists for more than 20 years and is used by most people on a regular daily basis nowadays, we expect some forms of dialect spellings as being established and being preferred over others (Tophinke 2008), even though we cannot observe a standardisation of spelling as reported on other non-standard spellings (Heyd 2016). For researchers interested in non-standard spellings, CMC is a gold mine for investigating individual writing habits as well as group-specific practices.

For researchers using corpora to investigate CMC, non-standard spellings are only one aspect of “noisy data”, and therefore one of the big challenges for automatic processing of CMC data. However, in this paper, we will not focus on problems for natural language processing caused by non-standard spellings. We will concentrate on non-standard spellings used on Social Network Sites from the perspective of variational linguistics. Furthermore, we are interested in extralinguistic features of the authors of such messages. We will therefore consider sociolinguistic relevant information in the descriptions and analysis that we provide in this article.

We will start with a short introduction to non-standard writing practices and spellings in general (Section 2) and continue with a case study in which we analyse selected spellings in the DiDi Corpus, a corpus that contains a high percentage of German dialect messages (Section 3). We will also introduce a comparison of the results from the DiDi Corpus to spellings in two other German CMC corpora, the Swiss SMS Corpus and the Dortmund Chat Corpus (Section 4) and summarize our results in Section 5.

2. Writing in Society

Writing is a socially relevant action; it is a social practice. Texts are usually designed for an audience and mean something. In order to be able to produce such texts, people must be literate, they need knowledge about how to write (and read). However, literacy is not limited to these skills. To be literate also means to apply these skills “for specific purposes in specific contexts of use” (Scribner & Cole 1981: 236), in order to communicate and gain information within a society. The term literacy comprises a lot of skills and competences. One of them, a very basic one, is the knowledge about orthography. Applying literacy skills also implies the adoption of orthographic knowledge.

Sebba claims that using orthography is by itself a practice and that it is strongly connected to specific literacy practices and societal attitudes towards them. Moreover, using orthography is even very ideological and is based on “beliefs about what language is, should be, and should be used for” (Sebba 2007: 25). Sebba differentiates between ‘legitimate’ and ‘illegitimate’ writing practices. In ‘legitimate’ writing practices (writing at school, publishing in newspapers and books), people will most likely adhere to orthographic rules. In
‘illegitimate’ writing practices, such as graffiti, and ‘marginal’ writing practices, such as text messages and online chats, people “have greater freedom to deviate from the norm” (Sebba 2007: 25). Variation is much more likely to occur in ‘illegitimate’ and ‘marginal’ writing practices.

Looking at deliberate variation of conventional norms makes the notion of orthography as practice better understandable. Deliberate variation means that people make a choice between variants. The mere possibility of variation is a precondition for generating social meaning with orthography (Sebba 2007: 32-33).

2.1. Non-standard Spellings

There is of course a difference between ‘licensed’ and ‘unlicensed’ variation. Licensed variation can either carry no meaning as in Example (1) in which we find two variants that are equal in meaning and both are possible in German orthography. Licensed variation can also bear extralinguistic meaning such a regional origin, as Example (2) shows. ‘Unlicensed’ variation is probably more conspicuous. It covers all variants that are not licensed by a given orthography (Example (3)).

1) nervös vs. nerves ‘nervous’
2) colour vs. color
3) motha vs. mother

A deviation generates social meaning only if it is recognisable as a variation of a familiar conventional form. Hence, the variation of a word cannot be completely arbitrary, it must make use of possible alternatives (cf. Dürscheid & Stark 2013: 194), which are constrained to a large extent by the conventional sound-letter correspondences of a particular orthographic system. One conclusion of this is that “even unlicensed deviation is constrained to a large extend by the conventional norms.” (Sebba 2007: 31). Another conclusion would be that everyone who wants to create socially relevant meaning through spelling needs advanced knowledge of the orthography of the respective language. Therefore, we can consider spellings as a matter of literacy.

The kind of meaning a spelling variant bears depends on the purposes of variation of the person or the social group that is applying such variations. There are many different types of non-standard spellings. Androutsopoulos (2000: 521-522) has categorised those types based on an analysis of German fanzine writings which has been taken up by many other researchers (e.g. Jaffe 2012). He distinguishes between:

1) phonetic spellings: representations of standard pronunciation not covered by standard orthography, e.g. German supa for super [ˈzuːpe] ‘great, excellent’;
2) colloquial spellings: reduction phenomena typical for colloquial speech;
3) regiolectal spellings: features typical of a regional variety;
4) prosodic spellings: representations of prosodic patterns, e.g. word stress in capital letters;
5) interlingual spellings: phonetic spellings of loanwords according to native orthographic rules, e.g. German Aktichn for English action;

---

4 The term marginal may be misleading in this context and refers to Sebba’s model of an “orthographic space” in which text types can be placed according to their status of regulation. The most regulated orthographic space is in the centre of the model and less regulated orthographic spaces at the margins where the pressure of regulation is less high (Sebba 2007: 43).
6) **homophone spellings**: graphic alternations without a correspondence to phonetic alternations, e.g. lexical substitutions such as *u* for *you*, *8* in *n8* for *night*; grapheme substitutions: *ph* for *f* in *phat*.

All the types of non-standard spellings have in common the fact that they are able to “place the discourse in which they appear in an opposition to standard language and the ideologies associated with it” (Androutsopoulos 2000: 515). This potential of non-standard spellings is not only important for subcultural writing practices but may become relevant also for CMC.

### 2.2. **Dialect Spellings**

What has been stated for non-standard spellings in general also holds true for regiolectal or dialectal spellings: occasions in which dialect spellings are used define them as being in opposition to formal language settings, i.e. settings in which standard language use is expected. With Androutsopoulos (2000), we differentiate between an indexical function of dialect spellings and a symbolic function. An indexical function becomes obvious when writers point to a certain group of people (e.g. farmers) in order to stigmatize them and attribute a lower social status to them. The symbolic function of dialect spellings, on the other hand, becomes evident when dialect speakers start to write the variety they use in speech in order to communicate in written form. With the rise of CMC, dialect spellings are not only selectively used by one person occasionally but can be observed in many communities. By this way, they are collectively adopted. To quote Androutsopoulos again: “When collectively adopted, such a practice can gain the status of an anti-standard (Halliday 1979) which functions as an important expression of social and cultural identity” (Androutsopoulos 2000: 515).

We will now turn to the strategies adapted for dialect spellings. Lötscher (1989) has systematised different strategies of dialectal spellings based on Swiss German Dialect (SGD) prose. He calls the collective dialectal spellings in a text a “grapholect” and describes diverse strategies of adaption of the regional variety based on the respective standard orthography (Lötscher 1989: 276-279):

- **A direct transfer** from standard orthography is the easiest and most common strategy. Graphemes of standard orthography are adopted with the corresponding phonetic value to spell a dialect word. Everyone capable of standard orthography can read and pronounce it properly.
- **A systematic reinterpretation** in contrast is very rare. One grapheme (e.g. *<k>* in Standard German orthography, pronounced as *[kʰ]*) will be used to represent a different sound (e.g. *[kʃ]* in SGD which does not exist in Standard German pronunciation) without indicating this reinterpretation.
- **The use of special letters** that are added to the conventional inventory is also very rare, one example is the use of *<y>* for long and short *[i]* (vs. *<i>* for *[i]*) in SGD.
- **Neutralisations of phonetic differences** of the dialect is often the case. Usually, the inventory of graphemes is too small to represent all dialect sounds. This is also typical for many standard spellings (e.g. the use of *<i>* for *[i]* and *[i]* in Standard German).

There are of course problems and challenges for those who want to write dialect, e.g. the perception of the sounds can be individually different and can thus lead to different spellings. This fact could be one reason for incoherent spellings in dialect texts. In addition, standard orthography often allows for variation, which will be transferred to dialect spellings. Adaptations can therefore be variable depending on individual preferences and choices.

However, because it seems unlikely that different writers use the same dialectal spellings (i.e. one grapholect), there are groups of people trying to specify rules for dialect
spellings in special use cases, e.g. Wikipedia entries, where some conventions for some dialect spellings are already proposed.\(^5\)

There are quite a lot of studies investigating dialect spellings in CMC. In studies on German CMC, it is often reported that dialect use in CMC is more frequent in the south than in the north of the German speaking area (Siebenhär 2006). In particular, Switzerland stands out in such comparisons. This is not surprising since dialect use in oral communication has a high local value in Swiss public life which is not the case in Germany, in particular in the north (Projektruppe Sprachenstellung 2009). Interestingly, recent publications on dialectal spellings in Swiss SMS (e.g. Felder 2015) rather stress the fact that dialectal spellings are not that deviant from standard orthography as one could expect, but people adhere in many cases to the standard. Dialect spellings in CMC are usually not radical but readable adaptions of standard orthography.

Another finding that is often confirmed is the fact that younger writers use dialect spellings more frequently than older ones. This was for example one finding of Peersman et al.’s (2016) study on Flemish SNS. The authors also found regional differences regarding the frequency of dialect spellings: in the west of Flanders dialect spellings are more common.

Further important aspects of dialect spellings can be found in Eisenstein’s (2013, 2015) research on American English Tweets. Unlike the aforementioned studies that focus on the word level, Eisenstein shows that non-standard spellings are sensitive to phonetic differences that go beyond word level. Phenomena such as *t deletion* at the end of words like *just* depend on the subsequent sound. The deletion is impeded when the next word starts with a vowel. The phonetic environment therefore must be considered. However, an important note comes from Childs’ (2016) study on an Afro-American community in the Appalachian Mountains who do not use the characteristics of their regional speech in written instant messages but features of Afro-American English, which shows that spellings do not necessarily reflect speech. In this case, the community chose spellings that highlight different aspects of their identity (e.g. ethnicity) and do not focus on regional aspects.

In summary, it is important to stress that frequency of dialect spellings may vary regionally and may not be found in all regions. In addition, age is an important factor for the occurrence of dialect in CMC. Finally, deviations can tell much about the writers’ pronunciation nevertheless one should be careful with the interpretation since spellings reflect only a choice of one’s identity as Child’s (2016) study has shown.

3. Non-Standard Spellings in South Tyrolean SNS

3.1. Languages in South Tyrol

Before we start analysing South Tyrolean CMC data, we will provide some basic information about the linguistic situation in the province of Bolzano, the northernmost province of Italy. Three languages are recognized as official languages in the territory – German, Italian and, in some valleys (Val Gardena and Val Badia), Ladin.

According to the 2011 Census (Autonome Provinz Bozen – Südtirol 2015), 69.4% of South Tyrolean inhabitants claim to belong to the German language group, 26.1% to the Italian and 4.4% to the Ladin language group. However, language groups are not equally distributed over the territory. Figure 1 shows the distribution of inhabitants according to the majority of the language group per municipality: In most parts of South Tyrol, the majority of speakers belongs to the German language group. Exceptions are the city of Bolzano and some

\(^5\) For example, for Scottish: [https://sco.wikipedia.org/wiki/Wikipedia:Spelling_an_grammar](https://sco.wikipedia.org/wiki/Wikipedia:Spelling_an_grammar) and Bavarian [https://bar.wikipedia.org/wiki/Wikipedia:Wia_schreib_i_a_guads_Boarisch%3F](https://bar.wikipedia.org/wiki/Wikipedia:Wia_schreib_i_a_guads_Boarisch%3F) and some other dialects.
municipalities in the South of Bolzano where the Italian language group is the majority, the valleys Val Gardena and Val Badia where the Ladin language group is dominant, and the city of Merano where the German and Italian language group are almost equally represented.

Sprachgruppenzugehörigkeit (in %) (Volkszählung 2011)

Figure 1 does not however take into consideration the fact that two varieties of German are used in South Tyrol. Standard German (SG) can be found in written documents, in radio and TV broadcasts, and is the language of education both at schools and at the University. It is also used in interactions with German speaking visitors from abroad. In interactions among German-speaking South Tyroleans, South Tyrolean Dialects (STD) are preferred. All social groups, no matter the social status, use STD in almost all occasions of everyday life (e.g. in family, at work; except the ones mentioned before). Among the German speaking population, STD has a high value (Lanthaler 2007: 230-231) and can be found in oral and written interactions in CMC. Example (4) illustrates the use of STD in CMC.

(4) STD: weil i es in dr schual brauch
SG: weil ich es in der Schule brauche
EN: because I need it in school

STD: lei wenn du es gor nimmer brauchsch ;D
SG: nur wenn du es gar nicht mehr brauchst ;D
EN: only if you do not need it anymore ;D

Example (4) is an authentic Facebook chat message of an 18-year-old girl from Merano. In this short message, we can observe different kinds of dialect spellings, which can either refer to

- differences regarding the sound/vowel systems of the dialect and standard: STD schual [ʃuːl] vs. SG Schule [ʃuː.lə] ‘school’, STD gor [ɡəɾ] vs. SG gar [ɡar];
- r-vocalisation or inhibited r-vocalisation: STD dr [dʀ] vs. der [dɐ] ‘the’;

- apocope: STD i [i:] vs. SG ich [ɪʃ] ‘I’;
- differences regarding the inflection suffixes of verbs: STD brauch vs. SG brauche ‘1. pers. sg. of brauchen ‘need’’, STD brauchsch vs. SG brauchst ‘2. pers. sg. of brauchen ‘need’’;
- or differences in the vocabulary: STD lei vs. SG nur ‘only’.

The sequence in Example (4) was taken from the DiDi Corpus which we will introduce in the next section.

### 3.2. The DiDi Corpus

The DiDi Corpus of South Tyrolean CMC data is a multilingual language corpus (cf. Frey et al. 2016). It consists of around 600,000 tokens collected from 136 profiles of Facebook users residing in South Tyrol, Italy (cf. Table 1).

<table>
<thead>
<tr>
<th>Table 1: The DiDi Corpus (Frey et al. 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>profiles</td>
</tr>
<tr>
<td>texts</td>
</tr>
<tr>
<td>tokens</td>
</tr>
</tbody>
</table>

In the DiDi Corpus, several aspects are annotated on token and message level. Languages were identified on message level using the language identification tool lang.py (Lui & Baldwin 2012). All messages were then tokenized with the twitter tokenizer ark-tokensize.py (https://github.com/myleott/ark-tokenize-py) and a subsequent manual correction of errors. A language specific part-of-speech tagging and lemmatization was necessary to handle different languages in the corpus. The RDR part-of-speech tagger (Nguyen et al. 2014) was used for Italian messages and the TreeTagger (Schmidt 1994) for German, English, Spanish, French and Portuguese messages. In addition, German messages were manually normalised, i.e. the assumed equivalent of Standard German was added following standard orthography for all non-standard writings (dialect as well as typos and other deviations). Based on the normalisations, German messages were automatically classified as dialect or non-dialect messages, allowing for a third class of not-classifiable messages (e.g. short and ambiguous messages). Moreover, the corpus contains annotations for typical CMC phenomena such as emoticons, emojis, hashtags, links and iterations of characters. Table 2 gives an overview of the distribution of the most used languages in the DiDi Corpus. The main languages of the corpus are German, Italian and English.

<table>
<thead>
<tr>
<th>Table 2: Distribution of languages in the DiDi Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>Italian</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>other</td>
</tr>
</tbody>
</table>

Table 3 shows the percentages of dialect messages in the German messages of the DiDi Corpus.

<table>
<thead>
<tr>
<th>Table 3: Percentage of dialect messages in German text messages in the DiDi Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>dialect</td>
</tr>
<tr>
<td>German</td>
</tr>
</tbody>
</table>
In addition, the corpus provides a series of metadata about the authors of the messages such as first language (L1), dialect speaker, valley of the dialect, age, sex, occupation and education which were obtained through an online questionnaire and which can now be used for sociolinguistic research. Glaznieks & Frey (2018) have analysed, among other things, quantitative aspects of dialect use by age. The results show that SNS users under 20 years of age use more dialect than any other age group. People between 40 and 50 years usually write fewer dialect messages. Figures 2 and 3 provide the percentages of dialect and non-dialect messages by age for status messages and comments.

![Figure 2: Percentage of dialect messages in German status messages by age in the DiDi Corpus (cf. Glaznieks & Frey 2018: 871)](image1)

![Figure 3: Percentage of dialect messages in German comments by age in the DiDi Corpus (cf. Glaznieks & Frey 2018: 871)](image2)

As the DiDi Corpus comprises a considerable percentage of dialect messages, it can be used for further qualitative analyses of non-standard spellings. In doing so, we have three main questions in mind that we would like to address with the help of the corpus data:
1) Are there common patterns in non-standard spellings?
2) Is there variation in non-standard spellings?
3) Are there extra-linguistic features that determine variation in non-standard spellings?

3.3. **Corpus Analysis: Non-Standard Spellings of Unstressed SG <-er>**

We will address the research questions raised in the previous section in a case study of non-standard spellings corresponding to unstressed word-final SG <-er> in polysyllabic words, exemplifying our analysis with authentic spellings from the DiDi Corpus. Unstressed word-final <-er> has been investigated to be subject to variation in other studies on German CMC (e.g. Tophinke 2008, Huber & Schwarz 2017) as well as on English SMS (e.g. Childs 2016). Variation is possible for standard phonetic spellings, e.g. super instead of super ‘great, excellent’ (Androukopoulos 2000: 521), and regiolectal spellings (see below); thus, this position is expected to be prone to variation also in the DiDi Corpus.

3.3.1. Method

As described above, the data in the DiDi Corpus contain a normalisation layer, which provides SG spellings of the original words in German SNS messages. Benefitting from this normalisation, it is possible to search for SG spellings on this layer no matter how the respective words are spelled in the original messages.

We thus extracted a list of words containing all words ending in <-er> on the normalisation layer, but limited the extraction to all messages recognised as German by the automatic language identification tool langid.py (Lui & Baldwin 2012). As the research focus was on unstressed syllables, we were looking for polysyllabic words in which the word stress is on a different syllable. Therefore, we manually excluded monosyllabic words because they (might) behave phonologically different from unstressed syllables in polysyllabic words, which may also influence the spelling (e.g. der [der] ‘the’ vs. wieder [ˈviːdɐ] ‘again’). In addition, we excluded a number of false positives within the list of polysyllabic words, e.g. words ending in <-ier> (e.g. Kurier ‘courier’), and compounds with a monosyllabic primary word (e.g. daher ‘therefore, therefrom’ or Wolkenmeer ‘sea of clouds’). Finally, we excluded all abbreviations on the original message layer such as o for oder ‘or’ as they do not give any insight in the spelling.

3.3.2. Results

In a second step, we categorised all spellings according to the used letters in the original message. Table 4 provides an overview of the data set used for the analysis.

<table>
<thead>
<tr>
<th></th>
<th>&lt;-or&gt;</th>
<th>&lt;-o&gt;</th>
<th>&lt;-er&gt;</th>
<th>&lt;-e&gt;</th>
<th>&lt;-ar&gt;</th>
<th>&lt;-a&gt;</th>
<th>&lt;-r&gt;</th>
<th>rest</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>337</td>
<td>499</td>
<td>16,980</td>
<td>44</td>
<td>99</td>
<td>252</td>
<td>1,600</td>
<td>16</td>
<td>19,827</td>
</tr>
<tr>
<td>excluded</td>
<td>149</td>
<td>130</td>
<td>5,042</td>
<td>31</td>
<td>85</td>
<td>112</td>
<td>882</td>
<td>13</td>
<td>6,444</td>
</tr>
<tr>
<td>data set</td>
<td>188</td>
<td>369</td>
<td>11,938</td>
<td>13</td>
<td>14</td>
<td>140</td>
<td>718</td>
<td>3</td>
<td>13,383</td>
</tr>
</tbody>
</table>

The vast majority of spellings in the DiDi Corpus matches SG <-er> (89.2% of all words), with only 10.8% deviating from it (cf. research question 1). In the following, we will
only focus on and refer to the latter as non-standard spellings since <-er>-spellings cannot clearly be assigned to a certain variety, be it SG or a non-standard German variety (cf. Glück & Glaznieks in prep. for a more detailed discussion). On the other hand, Table 4 also shows spelling variation, revealing seven variants: <-or>, -o, -er, -e, -ar, -a, -r>. The biggest group of variants is <-r> (49.7% of the non<-er>-variants) followed by <-o> (25.6%), <-or> (13%) and <-a> (10%) (cf. research question 2).

Table 5: Data set by categories and valleys (absolute numbers)

<table>
<thead>
<tr>
<th></th>
<th>&lt;or&gt;</th>
<th>&lt;-o&gt;</th>
<th>&lt;-er&gt;</th>
<th>&lt;-e&gt;</th>
<th>&lt;-ar&gt;</th>
<th>&lt;-a&gt;</th>
<th>&lt;-r&gt;</th>
<th>rest</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolzano</td>
<td>79</td>
<td>12</td>
<td>4,802</td>
<td>5</td>
<td>1</td>
<td>27</td>
<td>530</td>
<td>2</td>
<td>5,458</td>
</tr>
<tr>
<td>Eisack</td>
<td>10</td>
<td>37</td>
<td>1,100</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>1,168</td>
<td></td>
</tr>
<tr>
<td>Valley</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merano</td>
<td>52</td>
<td>0</td>
<td>3,419</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>68</td>
<td>0</td>
<td>3,555</td>
</tr>
<tr>
<td>Puster</td>
<td>0</td>
<td>314</td>
<td>616</td>
<td>4</td>
<td>90</td>
<td>2</td>
<td>1</td>
<td>1,027</td>
<td></td>
</tr>
<tr>
<td>Valley</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sarntal</td>
<td>1</td>
<td>0</td>
<td>475</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>31</td>
<td>0</td>
<td>513</td>
</tr>
<tr>
<td>Übergresch-Unterland</td>
<td>46</td>
<td>0</td>
<td>699</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>765</td>
<td></td>
</tr>
<tr>
<td>Valley</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vinschgau Valley</td>
<td>0</td>
<td>0</td>
<td>180</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>55</td>
<td>0</td>
<td>238</td>
</tr>
<tr>
<td>no dialect</td>
<td>0</td>
<td>6</td>
<td>597</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>607</td>
</tr>
<tr>
<td>no data</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>188</td>
<td>369</td>
<td>11,938</td>
<td>13</td>
<td>14</td>
<td>140</td>
<td>718</td>
<td>3</td>
<td>13,383</td>
</tr>
</tbody>
</table>

Table 5 shows that ‘consonantal spellings’, containing <-r> with or without an additional vowel, are preferred in the western part of South Tyrol (primarily Vinschgau Valley, Merano and Bolzano), while ‘vocalic spellings’ lacking <-r> are preferred in the eastern part (Puster Valley). Furthermore, certain spellings occur more often in some valleys and do not occur at all in others (cf. research question 3). For example, <-o> and <-a> occur very often in the Puster Valley but not at all or in a very low frequency in the Vinschgau Valley. The Vinschgau Valley on the other hand shows more <-r>-spellings which in turn occur in the Puster Valley with a very low frequency. To test the significance of the distribution in Table 5, we conducted a Pearson’s Chi-Squared test which revealed that the distribution by valley is significant ($\chi^2 = 4701.5$, df = 64, p-value < 2.2e-16). However, the valleys in the centre of the territory, between the westernmost Vinschgau Valley and the easternmost Puster Valley, do not show a clear preference for a certain alternative spelling. An exception is the city of Bolzano with a preference for <-r>-spellings and the southernmost region Überetsch-Unterland with a preference for <-or>-spellings. To answer research question 3, we can identify at least one extra-linguistic feature – the geographical origin of the writers – that seems to be relevant for variation and consistency in spellings.

As age is also one of the metadata available in the DiDi Corpus, we analysed the distribution of spellings by this variable.

Table 6: Data set by categories and age (absolute numbers)

<table>
<thead>
<tr>
<th></th>
<th>&lt;or&gt;</th>
<th>&lt;-o&gt;</th>
<th>&lt;-er&gt;</th>
<th>&lt;-e&gt;</th>
<th>&lt;-ar&gt;</th>
<th>&lt;-a&gt;</th>
<th>&lt;-r&gt;</th>
<th>rest</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;20</td>
<td>93</td>
<td>7</td>
<td>625</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>535</td>
<td>0</td>
<td>1,272</td>
</tr>
<tr>
<td>20-29</td>
<td>51</td>
<td>227</td>
<td>1,568</td>
<td>4</td>
<td>6</td>
<td>75</td>
<td>80</td>
<td>1</td>
<td>2,012</td>
</tr>
<tr>
<td>30-39</td>
<td>1</td>
<td>26</td>
<td>2,951</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>59</td>
<td>2</td>
<td>3,052</td>
</tr>
<tr>
<td>40-49</td>
<td>0</td>
<td>6</td>
<td>2,064</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>2,096</td>
</tr>
<tr>
<td>50-59</td>
<td>43</td>
<td>49</td>
<td>4,125</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>40</td>
<td>0</td>
<td>4,273</td>
</tr>
</tbody>
</table>
The distribution of spellings in Table 6 shows that younger users (under 30 years old) in the DiDi Corpus prefer <-o>-spellings over <-a>-spellings. However, both variants are characteristic for younger writers. Please note that consonantal <-r>-spellings are also very characteristic for younger writers. The Pearson’s Chi-squared test ($\chi^2$-squared = 5277.7, df = 35, p-value < 2.2e-16) again revealed that this distribution is significant.

To sum up the results of the analysis, the data show an overall preference for <-er>-spellings (cf. research question 1) as well as spelling variation (cf. research question 2). We have found seven variants in the data. However, considering extralinguistic variables (cf. research question 3), such as geographical origin or age of the writers, again common patterns (cf. research question 1) within groups of writers become visible.7

### 3.4. Non-standard Spellings of Unstressed SG <-er> and Phonetics in South Tyrol

In the previous section, we have shown that at least two extra-linguistic variables influence non-standard spellings in South Tyrol: the geographic origin and the age of the writers. In the present section, we will discuss to what extent these spellings match the phonetic realisation.

With regard to the origin of the writers, ‘vocalic spellings’ are favoured in the eastern part of South Tyrol while ‘consonantal spellings’ dominantly occur in the central and western part. These findings are in line with two dialect atlases from different times, with Klein & Schmitt’s data from the middle of the 19th century (1965-1969) and with Scheutz’s data from the beginning of the 20th century (2016). Preferences for <-o>- and to a lesser extent <-a>-spellings in the Puster Valley even perfectly fit the [-o]-transcriptions (Klein & Schmitt 1965-1969) and [-o]/[-a]-transcriptions (Scheutz 2016) in both publications. Preferences for <-r>-spellings in the Vinschgau Valley are at least in line with the [-r]-transcriptions in Klein & Schmitt (1965-1969), while Scheutz’s (2016) transcriptions show an additional vowel before [r] for this area: [-or].

With regard to the age of the writers, <-o>- and <-r>-spellings are preferred by younger writers (under 30 years old). In this context, Scheutz’s (2016) data is of special interest. His study compares elder (above 70 years old) and younger (between 20 and 30 years old) people’s pronunciation, in this case even using the same words as Klein & Schmitt’s (1965-1969) study. With respect to the dialectal realisations of unstressed SG <-er>, the author finds a substantial difference between the age groups in the Puster Valley: younger informants favour [a] over [o] whereas [o]-pronunciations are well documented for elder informants. According to Scheutz, there is a substantial sound change observable, changing the highly characteristic [o]-sound to a less conspicuous [a]-sound. This contradicts the preference for <-o>-spellings by younger writers in the DiDi Corpus.

One reason why <-o>-spellings are favoured over <-a>-spellings by younger users, could be that <-o> is characteristic for the Puster Valley and therefore can perfectly function as a “graphemic contextualisation cue” (Androustopoulos 2000: 517) which explicitly signals a user’s origin, even though it does not match the actual pronunciation.

---

7 Unfortunately, it is not possible to do further analysis with mixed effect models since the number of users that has provided data to the DiDi Corpus is too small to form further subgroups. Testing the distribution for other metadata information, e.g. sex or message type is of course an option and could be done in further research.
4. Non-Standard Spellings of Unstressed SG <er> in other German CMC Corpora

In this section, we will compare the results from the DiDi Corpus to other corpora. There is only a small number of CMC corpora currently publicly accessible (January 2018) which limits the possibilities of any comparative analysis.

The first corpus we found for comparison is the Standard German and Swiss-German part of the multilingual Swiss SMS Corpus (Stark et al. 2009-2014), which contains approximately 18,000 SMS (around 464,000 tokens) collected from Swiss mobile phone users. The Swiss SMS Corpus is particularly interesting for comparisons regarding regiolectal spellings, as dialect use in German-speaking Switzerland is common in oral and written communication (Siebenhaar 2006).

The second corpus we found for comparison is the Dortmund Chat Corpus (Beißwenger 2013) version 2.1, part 1 of 2, which contains chats (around 625,000 tokens) from open chat platforms.8

We restricted the comparison study to the three most frequent polysyllabic words in the DiDi Corpus ending in unstressed <er> in SG aber ‘but’, wieder ‘again’ and oder ‘or’. Figure 4 shows the distribution of spellings in the DiDi Corpus for the respective words separated for dialect and non-dialect messages (according to automatic assignment, cf. Section 3.2). In messages recognized as dialect messages, the percentage of non-<er>-spellings is much higher than in messages recognized as non-dialectal in which we hardly find instances of alternative spellings to SG <er> in this particular position. Furthermore, we can see three frequent spelling alternatives to <er> – <r, -or, -o> – while occurrences of <a> and <e> are very rare (less than 1% of all spellings).

In the Swiss SMS Corpus, we found a similar proportion of spellings in SG messages compared to dialect messages: <er>-spelling is preferred over all other variants (less than 1%) in the SG messages whereas there is considerable variation observable in the dialect messages9 (cf. Figure 5). The preferred alternative spelling in dialect messages is <r> although there are also some other variants – <-a, -o, -e, -är> – with a very low frequency. With respect to this position, Swiss spellings in dialect messages are more homogenous compared to South Tyrolean spellings.

---

8 This version is accessible since 2017 without any restrictions via the CLARIN website: https://vlo.clarin.eu/record?2&docId=http://58.47.47.hdl.handle.net/47.11858.47.00-203Z-0000-002D-ECC7-2.

9 There are distinct subcorpora for Standard German and dialect messages available in the Swiss SMS Corpus, cf. https://sms.linguistik.uzh.ch/
In the Dortmund Chat Corpus, texts were not classified for dialect use. Figure 6 therefore shows the results for the entire subcorpus. The main finding is that there is almost no variation on that position although some instances of <-a>, <-r> and <-e> occur (less than 1%). Compared to the Swiss and the South Tyrolean data less variation is observable. In addition, it
seems that even phonetic spellings like <-a> are avoided and people stick to the orthographic rules.

Accepting that the three corpora compared contain different types of texts (Facebook messages vs. SMS vs. open platform chat messages), the comparison study reveals interesting differences and commonalities between the three corpora. Non-standard spellings are very common in South Tyrolean and Swiss texts and result in common spelling alternatives. However, they are not used in SG texts in CMC, nor in the DiDi Corpus, or in the Swiss SMS Corpus, or in the Dortmund Chat Corpus. On the other hand, standard orthographic spelling remains the preferred variant in dialect messages which indicates that it is optional to mark dialect on the analysed position, but is not essential for dialect texts.

5. Conclusions

In this article, we have analysed non-standard spellings on South Tyrolean SNS. We have introduced a case study on variants of SG <-er>-spellings in word-final unstressed syllables of polysyllabic words. The results of the case study identified two extralinguistic features that influence the choice of a non-standard variant, namely geographical origin and age of the writers. We have shown that the choice of non-standard spelling variants largely matches phonetic features of dialects spoken in the respective areas suggesting phonetic spelling in CMC. We have also identified one mismatch between written dialect data and spoken dialect data regarding preferences of younger SNS users, which we referred to as an instance of a graphemic contextualisation cue that is conventionally associated with a certain region. According to Androutsopoulos (2000: 517), *graphic contextualisation cues* “evoke certain frames of interpretation by establishing a contrast to the text’s spelling regularities or to the default spelling of a linguistic term”. Something more important than an exact phonetic spelling might be the stereotypical association of one region with a specific sound which is reproduced and thus retained in written dialect even though there is a sound chance observable in spoken dialect (Scheutz 2016: 55-58).

The results of our case study are not enough to exhaustively describe non-standard spellings for South Tyrol. Further analyses are necessary in order to complete the picture.
Promising results could be obtained from investigations of systematic phonetic differences between SG and STD as illustrated in Example (4). However, there are many problems and challenges linked to the interpretation of a non-standard spelling variant as an instance of a specific variety, e.g. a dialect, which we will discuss in detail in Glück & Glaznieks (in prep.).

BIBLIOGRAPHIC REFERENCES


PROJEKTGRUPPE SPRACHEINSTELLUNGEN (2009), Aktuelle Spracheinstellungen in Deutschland. Erste Ergebnisse einer bundesweiten Repräsentativumfrage, IDS and University of Mannheim.


