Crowdsourcing Swiss Dialect Transcriptions for Assessing Factors in Writing Variations

Simon Clematide[1], **Karina Frick**[2], **Noëmi Aepli**[1], **Jean-Philippe Goldman**[2]
[1]Institute of Computational Linguistics,
[2]Zurich Center for Linguistics,
University of Zurich,
simon.clematide@uzh.ch

Abstract

In this paper, we systematically analyze writing variations of Swiss German in two existing corpora with standard German glosses, a corpus of 10,000 short text messages and a corpus of transcribed oral history recordings (90,000 tokens). We show that neither resource is sufficient for assessing factors in writing variations of users and describe a data collection project involving a citizen science community for solving this problem. Laymen will independently and redundantly transcribe 1,200 short samples (15-20 seconds) of audio material in Swiss German according to their own best practice.

1 Introduction

Over the last two decades, with the rise of new media in our everyday lives, writing in Swiss German has become very popular and its usage has increased considerably in private written communication such as text messages, e-mails or Facebook postings (Siebenhaar, 2008, p.2). There can no longer be talk of a "medial diglossia" (Kolde, 1981, p.68), which assumes that spoken dialect and written Standard German are functionally divided. Other factors, such as formality, communicative immediacy and distance have become far more important regarding the choice between (written) dialect and Standard German. Moreover, the popularity of writing in dialect has a lot to do with the fact that no official standard norm exists for the orthography of the Swiss German dialect (Christen, 2004, p.77). That is to say that users writing in Swiss German cannot violate any norms or make any mistakes which could possibly be sanctioned; this might be one of the main reasons why many language users in the German-speaking part of Switzerland prefer using dialect in their private correspondence (Aschwanden, 2001, 62). Furthermore, dialect is connoted very positively for Swiss German speakers and is also regarded as emotional whereas High German is perceived as rather impersonal and aloof (Sieber, 2010, p.380).

2 Related Work

The non-existence of an orthographic norm leads to many different writing variants in private written communication, as, for example, Siebenhaar (2003; 2006) has shown for Swiss chat rooms. He finds that there is a great variety of dialect writings for 8 investigated lexemes (Siebenhaar, 2006, 233). Although there have been various efforts to unify the spelling of Swiss German dialects, e.g. by (Dieth, 1986) (1938) or (Marti, 1985) (1972), they do not have any influence on chat users. This is certainly owed to the simple fact that users normally do not know these expert guidelines because they are not taught in school (Siebenhaar, 2006, 54). Instead, as Siebenhaar (2003, p.134) points out, the written dialect observable in chat rooms reflects a spontaneous vernacular spelling which is not bound to any standard rules but rather to phonetic distinctions in the different Swiss German dialects, e.g. Bernese or Zurich German. That is why in some cases the non-standardized vernacular writing "[...] still reflects the geo-linguistic distribution described in the linguistic atlas of German speaking Switzerland SDS (1962-1997) based on recordings of the 1940s and 1950s." (ibid: 125). Next to the phonetic influence, social variables and individual preferences concerning the scripting play an important role (ibid: 134).

3 Materials and Methods

There exist two larger corpora of Swiss German where spelling variation can be measured by comparing different realizations of written words with respect to normalized standard German glosses. The first one, SMS4science, is truly user-generated

German	*English*	*H*	*Swiss German Variants normalized to lowercase (Frequency)*
			SMS4Science
nächste	next	4.1	nächst(23), nächscht(16), nöchst(13), nögscht(11), nögst(6), next(6), nechscht(5), nägscht(4), negst(4), nögsch(4), negsht(4), näxt(4), nöchscht(3), negscht(3), nächsti(2), nächste(2), näxti(1), nächsht(1), nächschti(1), nägst(1), nöchschti(1), nechst(1), nöchschte(1), nöchsti(1), nächschte(1), nechsti(1), nàxt(1), nöxst(1), nögschd(1)
wochenende	weekend	3.2	wuchenend(36), wuchenänd(19), wucheänd(13), wucheend(11), wochenend(4), wucheendi(4), wochenende(3), wochenänd(3), we(2), wuchäänd(2), wocheänd(2), wuchaend(2), wuchenendi(2), wuchaändi(1), wocheendi(1), wuchenäd(1), wuchend(1), wochänend(1), wuchänänd(1), wuchanend(1)
vielleicht	maybe	3.3	vilicht(62), villicht(22), viellicht(16), vilich(11), velecht(9), filicht(8), vilech(4), velicht(3), velech(3), villich(3), vellecht(3), vielicht(3), filich(2), vielich(2), vellicht(2), viellech(1), filcht(1), vielech(1), vielleicht(1), vilivh(1), viellecht(1), vilecht(1), vellech(1), vilichd(1)
ich	I	1.3	ich(2896), i(1791), ech(115), ig(50), e(33), ih(17), iich(14), ni(9), ìch(5), ch(4), eg(3), ii(3), y(3), ici(2), hch(2), icg(1), ych(1), ìg(1), ìcg(1), iiich(1), isch(1), ibh(1), 'ch(1)
			Archimob
nachher	afterwards	4.2	nachher(13), ne(10), nòchethèèr(8), nòchhèèr(6), nòchher(5), nachhèèr(4), na(3), nòhèèr(3), nacher(3), naher(3), nòcher(2), no(2), nòher(2), näächer(2), nòchhèr(2), när(2), nachhär(1), nochhèèr(1), neecher(1), nä(1), nähär(1), nor(1), nochher(1), nahene(1), nòchether(1), nachhäär(1)
erdapfel	potato	2.3	hèrdöpfel(4), häärdöpfel(4), härdöpfel(3), härdepfu(1), hòrdöpfel(1), hò̀ò̀rdöpfel(1)
vielleicht	maybe	0.6	vilicht(66), vilich(4), viilicht(1), vilìcht(1), vìlìcht(1)
ich	I	1.0	ich(1157), iich(214), i(115), ch(3), ii(3), si(1)

Table 1: Writing variations in Swiss German short messages and expert transcriptions including their overall entropy (*H*)

content of short text messages originally written in Swiss German. Apart from the phonetic distinctions, we find all kinds of idiosyncratic spelling behaviour in this material, according to the "anything-goes" orthography (Dürscheid and Stark, 2013). The second corpus, ArchiMob, contains content that was transcribed from audio material by trained linguists. Therefore, the spelling variations should only reflect the phonemic distinctions that were in the focus for this corpus. In the next section, we contrast these two very different resources.

SMS4Science The Swiss SMS4Science Corpus[1] contains 10,706 short text messages that are mainly written in Swiss German. All messages were donated by volunteers who could also provide sociolinguistic and demographic metadata by filling out a questionnaire with topics such as gender, age, domicile, mother tongue, SMS use, or the use of T9.

As described in Ruef and Ueberwasser (2013), all messages were tokenized and an interlinear glossing in mostly standard German wordings (existing helvetisms were used as much as possible) was manually added. The glossing also split fused Swiss German words[2] and clitics (e.g. "chani" (*can I*) into their corresponding and orthographically correct equivalents ("kann ich"). The manually created glosses were then automatically processed by two different morpho-syntactic taggers, the TreeTagger (Schmid, 1995) assigning standard part-of-speech tags and the RFTagger (Schmid and Laws, 2008) assigning fine-grained morphological tags. The latter would allow to search for specific inflected words, for instance, a verb form in first person singular present tense. However, in order to keep the evaluation of both corpora comparable, we ignore the morphological features of SMS4Science.

For our evaluation on writing variations in short messages, we focus on words with single word glosses and ignore the phenomenon of dialectal or orthographical fusion of words. Using the ANNIS query interface to the Swiss German SMS4science subcorpus we searched for all words with a single gloss in standard German. For technical reasons[3],

[1] See sms4science.ch. Of total 25,947 messages, 41% are Swiss German, 28% Standard German, 18% French, 6% Italian, and 4% Romansh.

[2] Sometimes purely idiosyncratic orthography shows up, e.g. "ichdenkedudörfschsichermitfahre" (*I think you can surely ride with us*).

[3] Unfortunately, the SMS4science corpus cannot be downloaded in a suitable XML format. In order to exclude writing variations that originate from fused words, for instance, "chani" (*can I*) as a variation of "kann" (*can*), we had to restrict

the last token of each message could not be retrieved and from the total of 288,434 Swiss German tokens we could collect 249,029 (86%). Of these, 1,677 were manually marked as abbreviations and therefore excluded from our statistics.[4] To keep the results from SMS4science and Archimob comparable, we normalized the glosses and the word forms to lowercase. 49,591 glosses appear only once, leaving us with 197,761 tokens where we actually might observe writing variation.

We suggest to quantitatively measure the amount of variation in terms of the minimal amount of bits needed for encoding all variants, thus taking into account the number of different writings v, and also their relative frequency p_v:

$$H(V) = -\sum_{v \in V} p_v * log(p_v)$$

In a corpus with a strictly normalized orthography (and without any typo), each gloss would have an entropy of 0. If a writing variation is very rare compared to the others, the entropy will 'weight' the relative importance of this uncommon spelling accordingly. Table 1 illustrates spelling variations found in the SMS4science corpus. The word "nächste" (*next*) has the highest writing entropy (H=4.1) of all words.

Fig. 1 shows the overall distribution of entropy plotted against the frequency of glosses and illustrates the broad range of variations. This figure only reports about words that contain at least one alphabetic character. Out of 5,963 different types that fulfill this condition, 2,941 (49%) show no variation and 3,022 show at least 2.

ArchiMob The *ArchiMob Corpus*[5] (Samardzic et al., 2016) consists of 34 transcribed interviews (528,381 tokens) with Swiss citizens who witnessed the Second World War. The recordings are taken from the Archimob[6] oral history collections, which contain 555 videos, out of which 300 are in Swiss German.

The compilation of the ArchiMob Corpus started in 2004 and the three transcription phases extended over a period ten years. For the different phases, not only the tools but also the guidelines changed. The guidelines follow roughly the Dieth script (Dieth, 1986) but do not make use of all available phonemic distinctions. The grave accents in "*nòchethèèr*" ('*afterwards*') mark open vowels, more examples can be seen in Tab. 1. Because this distinction was dropped in later phases, we removed these grave markers for the data shown in Fig. 2.

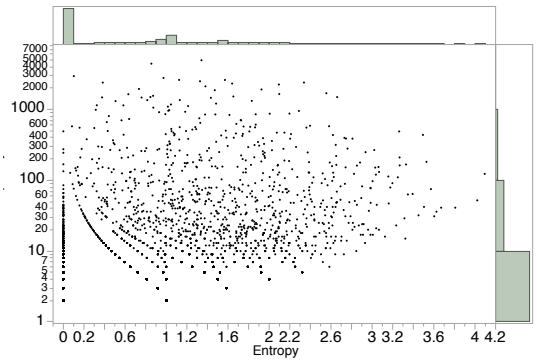

Figure 1: SMS4science corpus: plot of frequency of normalized words (y axis) against their writing variation entropy in Swiss German

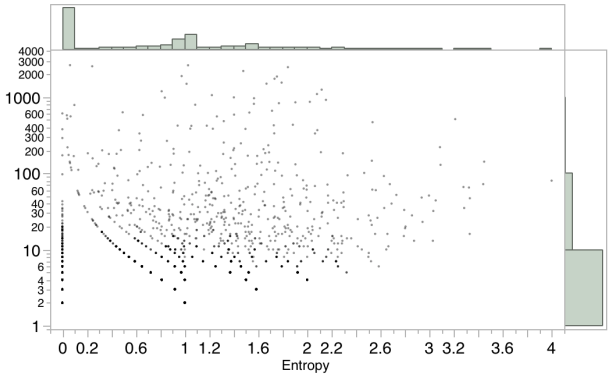

Figure 2: Archimob corpus: plot of frequency of normalized words (y axis) against their writing variation entropy in Swiss German

Furthermore, it has to be noted that not only the interviewees but also the transcribers have different dialectal backgrounds which, for instance, has an impact on the perception of vowels, leading to variations in transcriptions. Transcription variation has two sources: different dialects can use different words to refer to the same concept, and the same word can be pronounced and spelled differently. This results in a great number of potential variations which need to be reduced to a single canonical form in order to identify word variants. The general normalisation procedure is to transform every Swiss German word into the cor-

the query to tokens with a non-empty succeeding token.

[4] As can be seen in Tab. 1 in the row for "weekend", some abbreviations were not marked as such.

[5] www.spur.uzh.ch/en/departments/korpuslab/Research/-ArchiMob.html

[6] www.archimob.ch

Variations	*Alignment Output*
chaschmers sägä	`chasch-mers s-ägä`
chasch mirs sääge	`chasch mirs sääge`
can-you(-)me-it tell	Minimal Edit Distance = 4

Table 2: Pairwise Needleman/Wunsch sequence alignment of Swiss German transcriptions

responding standard German version following an etymological principle. Morphosyntactic features in Swiss German lexemes that are not implemented in standard German are transformed into morphologically transparent normalisations. For instance *dure* (through) does not exist in standard German, it would correspond to *durch* + *direction*, so it was normalised as *durchhin*.

At the current stage, only 6 recordings have normalizations attached to each word (93,455 tokens). For our evaluation, we dropped all fused words (2,915 tokens), which we identified by whitespace characters inside the normalization string. About 869 tokens did not have a valid normalization. For measuring the entropy, only words containing at least one alphabetic character were included. Out of 3,352 different types fulfilling this condition, 1,428 (43%) have no variation and 1,924 (57%) have at least 2 different spellings. Fig. 2 shows the observed spelling entropy, which in the case of ArchiMob should only express phonemic distinctions rather than personal writing and spelling habits.

Discussion Interestingly, Fig. 1 and 2 show a similar distribution although the underlying data was produced quite differently. These resources can be used for further explorations of typical pronunciation and writing variations in Swiss German. However, they both cannot be used to systematically correlate these variations with factors that might influence them. For the text messages, we are missing the phonetic form although we have real user-generated text. For the linguistic transcriptions, we are missing spelling variants, which native writers would produce. Therefore, we will collect new data in order to answer our research question.

Crowdsourcing Writing Variations The goal of our current project is to use a citizen science approach for collecting written Swiss German utterances as well as their standard German normalizations. Similar to the ArchiMob setup but different from the SMS4science setup, we will have spoken audio material that will be transcribed. However, the same material will be written in a spontaneous user-generated style (no guidelines, just the way they would write it in private communication) by several lay transcribers, which are to be recruited via a corresponding gaming platform on which users are able to locate Swiss German dialects with the help of the aforementioned audio stimuli.

These lay transcriptions give us the opportunity to assess the broad spectrum of spelling variations that is perceived as an adequate rendering of spoken Swiss German, and at the same time, correlate it with sociolinguistic factors that we assume to be relevant: (a) the dialect of the speaker and the transcriber (and their closeness), (b) the age and gender of the transcribers, (c) their expertise in writing in dialect. Accordingly, we are mainly interested in variation due to these social variables and not looking at variation caused by the medium or technical means, because we probably could not control the impact of the latter.

The consistency and variability of the independent parallel transcriptions can then be assessed automatically in a more fine-grained way. Character sequences can be aligned pairwise using sequence alignment algorithms (Needleman and Wunsch, 1970) as illustrated in Tab. 2.

We will also collect standard German "translations" of the Swiss German utterances, however, there will be no interlinear glossing in the style of SMS4science. Automatic normalization should be feasible given the available resources from SMS4science and ArchiMob, as shown in Samardzic et al. (2015; 2016).

User Interface Challenges for Transcription Transcribing audio recordings is a tedious and time-consuming task, especially for volunteering non-specialists. In the context of a web-based crowdsourced transcription project, volunteers should be extensively assisted in their transcribing task, or they would quickly give up. Usual facilitation for expert transcribers are all-in-one transcription software, or a USB pedal for convenient rewinding or slowing down of the speech rate, but none of them could apply here.

We will provide a simplified audio player with the usual facilities of playing and pausing as well as full and partial rewinding. Instead of displaying a continuous speech wave with a synchronized cursor

moving along the timeline, we represent the audio sample as consecutive blocks of speech segments. These speech units are pause-separated prosodic phrases, which corresponds to an average short-term memory span for audio transcription (Gentilucci and Cattaneo, 2005). As our audio material consists of about 1,200 15-to-20-second samples, the segmentation is automatically pre-computed with pause detection techniques[7] and should yield subsegments of 2-to-5 seconds for each sample. In the web interface, the user is able to play the full sample (with pausing at will) as well as to play segments individually. The current segment is highlighted. Eventually, simple keyboard shortcuts to avoid switching between keyboard and mouse are also available to enhance the user experience.

4 Conclusion

Systematically assessing factors of writing variation of Swiss German needs new resources that involve several transcriptions of the same audio stimulus. When dealing with highly user-specific writing habits, crowdsourcing transcriptions seems a natural approach for data collection. Independent transcriptions and their related sociolinguistic metadata enables us to investigate this phenomenon quantitatively. From an NLP perspective, acquiring more training material for automatic normalization of Swiss German is an important side effect.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant CRAGP1_164811/1 through the project "Citizen Linguistics: locate that dialect!" We would also like to thank the anonymous reviewer for his helpful comments on the first version of this paper.

References

[Aschwanden2001] Brigitte Aschwanden. 2001. »wär wot chätä?« zum sprachverhalten deutschschweizerischer chatter. online `http://www.mediensprache.net/networx/networx-24.pdf`.

[Christen2004] Helen Christen. 2004. Dialektschreiben oder sorry ech hassä text schribä. In *Alemannisch im Sprachvergleich. Beiträge zur 14. Arbeitstagung für alemannische Dialektologie in Männedorf (Zürich) vom 16.-18.9.2002*, ZDL-Beiheft 129, pages 71–85, Wiesbaden. Franz Steiner Verlag.

[Dieth1986] Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*. Lebendige Mundart. Sauerländer, Aarau etc. 2. Aufl. / bearb. und hrsg. von Christian Schmid-Cadalbert (1. Aufl. 1938).

[Dürscheid and Stark2013] Christa Dürscheid and Elisabeth Stark. 2013. Anything goes? sms, phonographisches schreiben und morphemkonstanz. In Martin Neef and Carmen Scherer, editors, *Die Schnittstelle von Morphologie und geschriebener Sprache*, Linguistische Arbeiten, pages 189–210. De Gruyter, Berlin.

[Gentilucci and Cattaneo2005] Maurizio Gentilucci and Luigi Cattaneo. 2005. Automatic audiovisual integration in speech perception. *Experimental Brain Research*, 167(1):66–75.

[Goldman2011] Jean-Philippe Goldman. 2011. Easyalign: an automatic phonetic alignment tool under praat. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 3233–3236, Florence, Italy.

[Kolde1981] Gottfried Kolde. 1981. *Sprachkontakte in gemischtsprachigen Städten. Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i.Ue.* Franz Steiner Verlag, Wiesbaden.

[Marti1985] Werner Marti. 1985. *Berndeutsch-Grammatik für die heutige Mundart zwischen Thun und Jura*. A. Francke, Bern.

[Needleman and Wunsch1970] S B Needleman and C D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53.

[Ruef and Ueberwasser2013] Beni Ruef and Simone Ueberwasser. 2013. The taming of a dialect: Interlinear glossing of swiss german text messages. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, volume 61-68 of *ZSM-Studien 5*. Shaker, Aachen.

[Samardzic et al.2015] Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of swiss german. In *Proceedings of the 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

[Samardzic et al.2016] Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2016. Archimob - a corpus of spoken swiss german. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4061–4066.

[7] Using tools like EasyAlign (Goldman, 2011) or WebMAUS (Strunk et al., 2014).

[Schmid and Laws2008] Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK.

[Schmid1995] Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the EACL SIGDAT-Workshop*. (überarbeitete Version).

[Siebenhaar2003] Beat Siebenhaar. 2003. Sprachgeographische aspekte der morphologie und verschriftung in schweizerdeutschen chats. *Linguistik Online*, 15(3):125–139.

[Siebenhaar2006] Beat Siebenhaar. 2006. Gibt es eine jugendspezifische varietätenwahl in schweizer chaträumen? In *Perspektiven der Jugendsprachforschung/Trends and Developements in Youth Language Research*, Sprache – Kommunikation – Kultur 3, pages 227–239. Lang, Frankfurt a.M.

[Siebenhaar2008] Beat Siebenhaar. 2008. Quantitative approaches to linguistic variation in irc: Implications for qualitative research. *Language@Internet*, 5(4).

[Sieber2010] Peter Sieber. 2010. Deutsch in der schweiz: Standard, regionale und dialektale variation. In *Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch*, HSK 35.1, pages 372–385. de Gruyter, Berlin, New York.

[Strunk et al.2014] Jan Strunk, Florian Schiel, and Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.