Swiss German Internet Language - Corpus Building and Analysis

Luc Aggett and **Dominic Fischer**
University of Zurich - Department of Computational Linguistics

Abstract

This paper describes the building process of a Swiss German Internet Language corpus including the associated difficulties and analyses key points of the use of language.

Comparing scraped comments from the reddit r/buenzli community against a formal written corpus of Swiss German, this work aims to highlight differences in vocabulary, high-level sentence structure by using Part of Speech-tags, as well as the challenges associated with collecting, annotating and analysing Swiss German language data.

1 Introduction

In the Swiss linguistic landscape, Swiss German stands out as a distinct and vibrant language variety. Studying Swiss German, however, poses unique challenges due to its informal, spoken nature and the limited availability of linguistic resources.

In recent years, the rise of online communities and social media platforms has had a profound impact on language use and variation. As one of the largest online communities dedicated to Swiss German language and culture, the subreddit r/buenzli offers a valuable source of Swiss German language data. By comparing scraped comments from r/buenzli against a formal written corpus of Swiss German, the NOAH's corpus consisting of newspaper articles from Blick, Swatch Investor Reports, Swiss German Literature and articles from the "Alemannische Wikipedia" , this study aims to shed light on differences in vocabulary and the distribution of Parts of Speech, highlighting the dynamic nature of Swiss German and its adaptation to the digital realm, and address the challenges associated with collecting, annotating, and analyzing Swiss German language data.

Language corpora have been widely used to investigate various linguistic phenomena across different languages and dialects, with the majority of existing corpora tending to focus on formal written language. In the context of Swiss German, the absence of universal rules and a written norm renders this impossible. Resources in Swiss German are scarce in a written form, and furthermore, due to the language's informal character, they might not retain a natural feel to them. The combination of these factors poses significant obstacles in understanding the lexical and grammatical features of this language variety.

By unraveling the vocabulary and Parts of Speech variations and challenges associated with Swiss German language data, this research advances our knowledge of language use and variation in online communities and highlights the importance of linguistic resources for preserving and studying regional dialects or low-resource languages such as Swiss German.

In the following sections, we will present the methodology employed in building the Swiss German Internet Language corpus, delve into the comparative analysis of formal written versus informal online language use, and discuss the results and implications.

2 Methodology

2.1 Data Collection

To gather the necessary data for this study, we utilized the Pushshift API to retrieve comments from the r/buenzli subreddit. It is important to note that the Pushshift API has been taken offline in the meantime. In order to keep the data availabla and ensure reproducibility, we downloaded the whole subreddit, resulting in a plethora of data. We employed python to obtain comments within specific time intervals defined by UNIX timestamps. The data collection process involved weekly downloads, resulting in almost 60'000 comments with more than 1.2 Million tokens in the form of JSON files spanning from May 23, 2013, to January 1,

2023. Each file contained comments for a particular week.

2.2 Data Cleaning

To ensure the quality and reliability of the collected data, we implemented a data cleaning process using python. This script performed various cleaning operations, including the removal of duplicate and deleted comments, as well as the elimination of for our purposes irrelevant meta-data. Additionally, we employed a simple language identification method based on the set of words used in the NOAH's corpus. By eliminating comments that did not contain any word also contained in that set of Swiss German words, we filtered out the most obvious non-Swiss German comments.

With that turning out not to be enough, another approach was devised, using specially trained language models to determine the language of a sentence. We tokenised the text into n-grams, contiguous sequences of n words. The tokeniser we used split the text into individual words and then generated n-grams from them. although n-gram based language models have been superseded by deep learning methods, due to our constraints in time, data and computing power we decided to use them nonetheless. We used simple count- and log-based probabilities for our models.

```
self.probs = {ngram: np.log(count[ngram]
            / total) for ngram in count}
```

We did not manage to implement smoothing for the model, so we instead went with a very low probability for unknown bigrams.

We created a predictor class that is responsible for training a n-gram language model based on n-grams. It takes a text and creates probabilities for each n-gram - in our case bigrams - based on its occurrence in the text. In doing so, the class builds a language model that reflects the statistical patterns of n-grams within a text, that is, a language. These probabilities serve as the foundation for predicting the probability of a given text being in the language the model was trained on.

We applied this class to create models for Swiss German, German, English, Dutch, French and Italian, using respective training data, and iterated over each comment. The Swiss German Sentences were quite well recognised, a feat which is hard to achieve with existing models. The non-Swiss German models did recognise the sentences in their respective languages, yet still contained a number of Swiss German sentences. We abandoned the Dutch model due to the corpus not containing a significant amount of dutch sentences. For the models for Italian, French and English, we intended to check and clean up their sentences using existing language detection libraries. We abandoned that idea due to unreliable results and the fact that the changes to our sentence selection would have been minor. In the end, the predicted language was added to each comment's metadata.

2.3 Tokenisation & Part of Speech-tagging

For the tokenisation of the r/buenzli corpus, we utilised a Swiss German tokenisation and Part of Speech (POS)-tagging model developed by Noemi Aeppli. This facilitated the segmentation of the corpus into individual tokens, enabling subsequent POS-tagging and analysis.

2.4 Data Analysis

To analyse the collected data, we utilised the programming languages Python as well as Rust - since the former did not suffice for certain of our needs - employing our own modules as well as existing libraries. With regards to the vocabulary, we had two focal points in mind, one being recognition of foreign words, the other word frequency and length analysis.

2.4.1 Language Identification

In our first attempts to identify foreign words within a given text, the employed methods relied on different python libraries. However, the most accurate and best-suited libraries were deprecated or no longer open source. After trying multiple libraries and approaches, even going into the source code of said libraries, we had to concede defeat. Due to the mixture of foreign words and not reliably detectable Swiss German words, the language detection libraries demonstrated suboptimal accuracy and yielded peculiar probability distributions for the language labels (e.g., ”de” having a probability extremely close to 1 for being Spanish, despite being the exact same word in French, Italian, Portuguese or Catalan).

Another approach centered around the utilization of bigrams extracted from NOAH's corpus. The idea was to derive a comprehensive set of Swiss German bigrams by analyzing the bigrams present in this corpus. This set would then serve as a reference for comparison with bigrams derived from the buenzli corpus. The comparison would result in a prediction score for each bigram, indicating its likelihood of belonging to Swiss German or being a foreign word, with a predetermined cutoff value making the decision. Any bigram below this threshold would be categorized as a foreign word, those above Swiss German words. The resulting set of foreign words would be further analysed and split into different languages using above mentioned language detection or a dictionary look-up. However, we had to abandon this approach due to the difficulty of creating the context windows. Using bigrams, each word would have had to have the preceding and the subsequent words saved, and this process exceeded reasonable time frames.

As a result, we devised a new approach to enhance the precision and reliability of language identification, specifically targeting the recognition of foreign words within Swiss German texts. We created sets from downloaded files containing an abundance of words in Italian, French and English. For each comment, we then checked whether any of its words was not in Swiss German but found in one of the thus created dictionaries. In that case, we appended the word to a dictionary together with the sentence that showcased its contextual usage.

2.4.2 Word Frequency Analysis

We intended to use our own implementation of the Damerau-Levenshtein distance algorithm for this task. This algorithm allows transposition as well as insertion, deletion and substitution at the cost of 1. Using it would enable fuzzy matching of words, and can therefore account for typos and dialectal variations. However, due to overly complex code, we settled for simple counting. An advantage of this is that typos or dialectal variants are still kept separate, which allows for more detailed analysis.

We opted to combined the Word Frequency Analysis with a POS-tag analysis, allowing us not only to see which words, but also which POS are used most often. Furthermore, it enabled us to gather information about the words' length and complexity.

Throughout the methodology, we prioritized data integrity and cleanliness. Our processes involved multiple stages of data collection, cleaning, tokenisation, and analysis. By employing a combination of our own Python scripts and existing modules, we strived to ensure accurate and meaningful insights from the collected data.

It is essential to acknowledge that this methodology has certain limitations, including the reliance on downloaded data due to the unavailability of the Pushshift API. Additionally, the language identification method employed in the data cleaning process may not capture all instances of Swiss German with absolute precision. Nonetheless, we took these limitations into account and proceeded with our analysis while being mindful of potential biases.

3 Results

3.1 Word Frequency Analysis

We decided to keep our Word Frequency Analysis simple and based our findings on the ten most frequent words of each POS-tag that we considered pertinent for this task.

The top adjective for both corpora turned out to be "guet" (engl. good). For the buenzli corpus, however, its frequency was almost three times higher. While potential negations before the adjective were not taken into account, together with the absence of any negatively connotated adjectives in the top ten, this might suggest a positive online environment. The presence of strongly opined adjectives (or misclassified adverbs) like "sicher" (definite), "richtig" (correct), "recht" (rather), "klar" (clear) as opposed to more neutral ones in NOAH's corpus ("neu" (new), "gross" (big), "lang" (long)) is in line with on of a forum's main purposes, making one's opinion known.

Nounwise, buenzli corpus featured two different spellings of a word used to denominate people ("lü(ü)t") as well as for the Swiss ("Schwizer"). Together with "land" (country) and "problem" (idem), key topics are centered around Swissness and Switzerland. NOAH's corpus uses exhibits different nouns used most frequently (yet no less Swiss), centering around time and Swiss products: "Tag" (day), "Uhr" (watch), "Kollektion" (collection), "Hotel" (idem), "Marke/Markä" (brand).

The range of the most frequent verbs was rather limited in both corpora, three different forms of "mache" (to do) occupying three of the top ten spots.

Looking at adverbs, very similar ones make it into the top ranks. Only one features in the buenzli but not in NOAH's corpus, namely "eifach" (just), likely hinting again at the fact that strong opinions are expressed.

Similarly, the particles used barely differ. Inter-

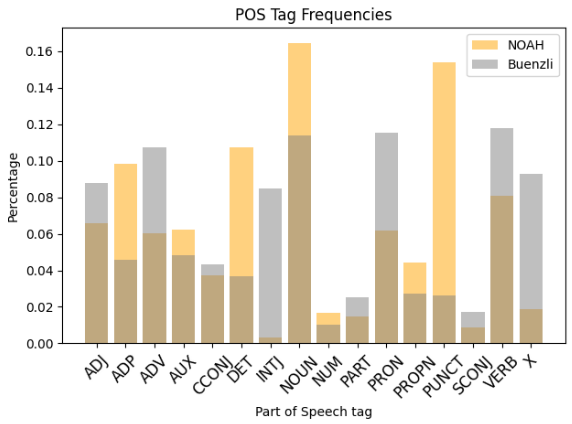

Figure 1: Part of Speech frequencies in the buenzli and NOAH's corpus

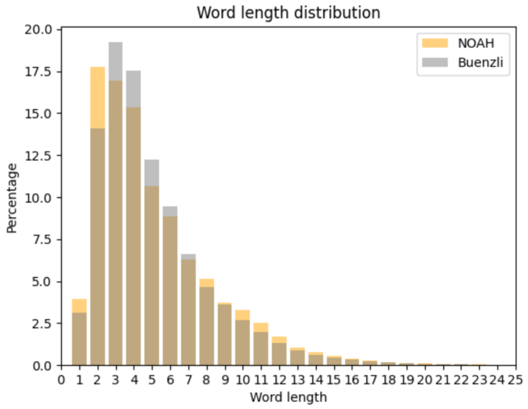

Figure 2: Word length distribution NOAH's / buenzli corpus

estingly, in both cases, we have three spellings for the negation particle ("nid/nöd/ned") with similar probabilities, showing the diversity in dialects and spellings.

3.1.1 Part of Speech-tag Analysis

Important observations can be made by not only comparing the actual words, but also the frequencies of their grammatical category in each corpus. Figure 1 visualises the differences. Buenzli exhibits less nouns than NOAH's corpus, but makes up for the deficit with an abundance of verbs - a tendency often observed in informal language. It seems reasonable, too, that informal speech tends to keep sentence structures less complex (lower percentage of subjunctions (SCONJ)), and would not have the same attention to punctuation as formal language.

Interjections, adjectives and adverbs, on the other hand, have a stronger presence in the buenzli corpus, displaying the expressiveness of informal speech. Similary, pronouns have a much stronger presence, while proper nouns are less frequent. This might indicate that the referents and topics are known within or inherent to the community, while external entities do not have the same importance as in NOAH's corpus.

Finally, adpositions and determines have much lower percentages in the online corpus, a tendency that manifests itself also in spoken Swiss German: if the direction or the noun is clearly specified as it is, adpositions and determines are more and more perceived as optional by speakers.

3.1.2 Word Length Analysis

The word frequency analysis also allowed us to see how big the differences were in terms of word length between the corpora. It turns out that the buenzli corpus' tended to use slightly longer words, the average being just below five characters for the buenzli corpus and marginally lower for NOAH's corpus. We think this is due to multiple factors: internet language uses abbreviations, emoticons, short forms et cetera, thus decreasing a token's average length. At the same time, there were very long tokens present that counterbalanced that: links to youtube videos, more or less random series of signs (such as a user demonstrating frustration, writing random characters and punctuation in a single string, or smiley faces using punctuation ("franzose..^^^^^^^^")), as well as other formattings that were not caught by our tokenising system ("&gt;armeehgwehr\n\ndas"). Furthermore, in informal written Swiss German, people tend to glue particles to words ("hanichem" instead of "han ich em") or make up new compound words ("kantonsdiskriminierend"), which have a tendency to become quite long.

3.2 Language Identification

Filtering out loan words was harder than we thought. Despite all the clean-up, identification was far from trivial. Lots of issues arose due to Swiss German's non standardised spelling. Firstly, basically any Swiss German word could be found in at least one of the other language's dictionaries. Secondly, loan words, if they have a significant frequency, for example in the context of youth language, they tend to get assimilated, becoming un-

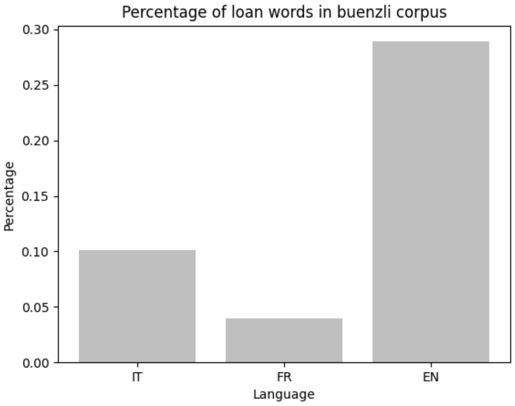

Figure 3: Foreign words percentages in buenzli corpus

recognisable in the process (e.g., “triggeret“ instead of “triggered“).

Looking through the data manually, we realised that many supposedly foreign words were not actually foreign words (e.g., “sell“: not English, but Swiss German for “should“). On the other hand, other words that should have been recognised, were not (e.g., “selle“, from English “to sell“, in Swiss German with the meaning of “to mess up“). We therefore think that while the data has to be taken with more than just a grain of salt, it still gives an idea as to what is to be found in the corpus.

It might be surprising how low the percentages of foreign words are. This is due to the fact that in this particular forum, Swiss German is really celebrated; the forum is meant to strengthen its status and the speaker’s linguistic awareness. Many speakers consciously avoided anglicisms, for example by using the literal translation “Pfoste“ to express the concept of the noun “post“ and, analogously, “pfostiere“ for the verb. Tendencies such as these go a long way to explain the limited use of foreign words.

It should be mentioned that many french words are very much part of the language already, to the point that they would be found in any dictionary. They did not feature in this statistic, explaining the low number.

Finally, despite all the clean-up we did, many more things would have needed to be undertaken in order to obtain reliable data. Named entity recognition, just to mention one measure, would have helped in distinguishing actual loan words from brand, people or place names.

4 Discussion

While this investigation provided valuable insights into the linguistic characteristics of Swiss German in different contexts, it is crucial to acknowledge and address the challenges encountered during the analysis process.

One major challenge stems from the nature of Swiss German itself. Swiss German is a highly diverse and non-standardized language with significant regional and individual variations. The absence of a standardized orthography and the prevalence of dialectal features pose considerable difficulties in establishing consistent criteria for corpus compilation and annotation. Consequently, variations in spelling, grammar, and vocabulary across the informal online context further complicate the analysis and interpretation of the data.

Moreover, the scarcity of resources specific to Swiss German further hindered the reliability of our results. The limited availability of linguistic references, lexical databases, and annotated corpora for Swiss German restricts the extent of linguistic analysis that can be conducted. For example, while Part of Speech-tagging is possible, it is still prone to errors, and morphological or syntactical analysis remain a vision for the future. As a result, our study relied heavily on manual annotation and quantitative analysis, which may introduce subjectivity and limit the depth of the findings.

Furthermore, the use of a Reddit forum as a source of Swiss German comments introduces additional challenges. Online forums are informal communication platforms where users often employ abbreviations, slang, and idiosyncratic expressions. Extracting meaningful linguistic patterns from such a dataset requires careful consideration and adaptation of analysis methods. The dynamic and rapidly evolving nature of online interactions also poses challenges in capturing a representative sample and accounting for potential biases in the data.

Given these limitations, it is important to interpret the findings of this study with caution. While our analysis provides initial observations and insights into the informal usage of Swiss German in online contexts, further research is needed to overcome these challenges and establish more reliable and comprehensive methodologies for analyzing Swiss German linguistic data.

While giving some insight into the differences in language, the comparison between the formal written Swiss German corpus and the Swiss German comments from the Reddit forum primarily highlights the complexities and difficulties associated with studying Swiss German in its inherently non-standardized contexts. Future research should aim to address these challenges, expand linguistic resources, and employ innovative approaches to advance our understanding of Swiss German variation and usage in diverse settings.

5 Conclusion

In conclusion, this study compared a corpus of formal written Swiss German to a corpus comprising Swiss German comments from a Reddit forum, aiming to investigate Part of Speech-tag and vocabulary variations including the presence of foreign words.

Significant differences were observed in the most frequent words apart from the category of function words. The variation in the frequency of POS-tags was found to be significant in all tags except for conjunctions. The use of foreign words constituted less than 0.45 percent of the total vocabulary, indicating minimal usage. Additionally, the analysis revealed that the word length in the online corpus was slightly longer compared to the formal written corpus.

It is important to note that the variation in the most frequent words does not enable us to draw secured conclusions. However, it does provided intriguing room for interpretation. The Buenzli corpus exhibited a higher frequency of positively connotated adjectives and expressive adverbs, as well as verbs and interjections, suggesting a positive online environment with strong opinions expressed. Adposition, determiners and punctuation did not have the same importance as in formal written Swiss German, possibly mimicking tendencies of informal spoken language. Nouns centered around Swissness and Switzerland, emphasising the forum's purpose of an intra-swiss exchange. These findings highlight the distinct characteristics of informal language use, with a focus on expressiveness and a tendency to simplify sentence structures.

Despite the observed differences in word length, the practical significance of this finding may be limited, with both means only marginally different just below of five characters per word. Nevertheless, it suggests that the informal online context may indeed make use of longer words, possibly influenced by factors such as the freedom of creating neologisms or the absence of strict writing conventions.

Regarding the presence of foreign words, the low proportion of foreign terms indicates that the Swiss German speakers of this particular forum predominantly rely on their native language in online interactions. This finding is consistent with the general trend of language preservation and maintenance within online communities dedicated to a specific linguistic variety, where individuals often prefer to use local or regional languages to express their identities and establish a sense of belonging.

Overall, this study contributes to the growing body of research on Swiss German by shedding light on the differences between formal written Swiss German and Swiss German comments from an online forum. It emphasizes the importance of considering the contextual factors and the influence of the online environment on language use. The findings underscore the resilience of Swiss German as the dominant language of online communication of Swiss German speakers, with minimal incorporation of foreign words, despite the increasing prevalence of English as a global language.

By gaining a better understanding of the linguistic variations and dynamics within Swiss German, researchers and language enthusiasts can contribute to the preservation and appreciation of this unique language variety, while also exploring its role in shaping cultural identity in the digital age.

Limitations

Despite the valuable insights gained from the comparison between the formal written Swiss German corpus and the Swiss German comments from the Reddit forum, it is important to acknowledge several limitations that may have influenced the study's outcomes and interpretation.

First, the sample size and specificity of both corpora may impact the generalizability of the findings. While efforts were made to collect a representative set of texts, the size and source of the corpora may not fully capture the diversity and richness of the Swiss German language as a whole. A larger and more diverse sample could provide a more comprehensive understanding of the language variations and patterns observed in different contexts.

Second, the selection of the Reddit forum as the source for informal Swiss German comments introduces potential sampling biases. The nature of online platforms makes it challenging to ensure a random and representative selection of comments. Furthermore, the specific subreddit chosen may have its own characteristics, attracting a particular subset of Swiss German speakers or topics of discussion. Consequently, caution should be exercised when generalizing the findings to all informal Swiss German online interactions.

Third, the selection of the NOAH's corpus as a reference corpus might introduce yet another set of issues. Due to those texts having been redacted artificially rather than coming into being from natural conversation, controversy about choosing them as a reference corpus cannot be avoided, despite the lack of other adequate resources.

Fourth and last, the reliance on manual or self-crafted analysis tools introduces subjectivity into the analysis process. Despite efforts to establish clear criteria and maintain consistency, the inherent interpretive nature of these methods may lead to variations in the identification and categorization of linguistic features. Utilizing automated or semi-automated annotation tools in future studies could enhance the objectivity and replicability of the analysis.