What's up, Switzerland?
A corpus-based research project in a multilingual country*

Simone Ueberwasser and Elisabeth Stark (Zürich)

Abstract

This paper offers some initial insights into the first large-scale and multilingual corpus of WhatsApp messages for linguistic research and the related research project "What's up, Switzerland?". Data was gathered in Switzerland in the summer of 2014 and will be made available to the academic public online at the end of the project (end of 2018). This article presents facts and figures about the corpus and the participants' demographic data as well as an overview of (the lack of) existing linguistic research in the field and the research intended in the SNSF-funded research project.

1   Introduction

This paper aims to present a recently begun research project on the language and use of WhatsApp messages in Switzerland, and first and foremost its common database, the first multilingual large-scale corpus of WhatsApp messages (617 chats, 763,650 messages, 5,543,692 tokens that can be used for linguistic research). Although the main device for mobile graphic[1] communication nowadays is WhatsApp, which has clearly replaced the older text messages (see Dürscheid/Frick (2014) on this issue), there were no sufficiently large databases available to investigate this new form of communication until our project was started. Contrary to the abundance of research on CMC in general and text messages in specific and contrary to the existence of a large number of CMC corpora (e. g. *sms4science Belgium:* www.sms4science.org; *sud4science Montpellier:* www.sud4science.org;, *sms4science Canada:* www.texto4science.ca; see also the table in Dürscheid/Stark (2011:303); the *Dortmunder Chat-Korpus:* www.chatkorpus.tu-dortmund.de; *IDS Wikipedia-Korpora:* www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html), WhatsApp communication is still a somewhat underresearched topic, something we would like to change, at least with regard to the situation in Switzerland where our data were collected and where the languages we are

---

[1] As *written* may sometimes refer to formal, elegant style, we prefer *graphic* as opposed to *phonic* when talking about the medial, material character of linguistic messages, i. e. based on letters (= graphic) or on sounds (= phonic). See Koch/Oesterreicher [2]2011 for a theoretical discussion of this important distinction, also on the terminological level, and Stark 2011 for its application to the analysis of text messages.

interested in (German, French, Italian, Romansh with their different dialectal varieties) are spoken. This paper is structured as follows: Section 2 presents a short overview of current research regarding WhatsApp messages; Section 3, the main part, presents the project's database, the Swiss WhatsApp corpus, with detailed information on how the data were gathered and prepared (sections 3.1 and 3.2), numbers and figures on the size of the corpus (number of messages, chats and emojis; section 3.3) and the demographics of the participants (age, gender, education, geographical origin, section 3.3.4); the final section, Section 4, concludes with a look forwards to the research to be undertaken on the data within the next two years.

2   State-of-the-art[2]

Since the beginning of this century, an important part of linguistic research has been dedicated to what is most frequently termed "computer-mediated communication", CMC. Important handbooks, monographs and seminal articles like Panckhurst (2007), Baron (2008), Thurlow/Mroczek (eds.) (2011), Herring/Stein/Virtanen (eds.) (2013), Thurlow/Poff (2013), Herring/Androutsopoulos (2015), and most recently, Dürscheid/Frick (2016) have analyzed multiple aspects of CMC, with a clear focus on graphematical, interactional and pragmatic aspects.[3] Swiss text messages, the very first popular form of mobile graphic communication, have been intensely investigated, as can be seen in the aforementioned books and also from the articles and monographs written in the context of the Swiss SNSF research project "SMS communication in Switzerland" (project number: CRSII1_136230): see, e. g., Bucher (2016), Cathomas (2015), Cathomas et al. (2015), Jucker/Dürscheid (2012), Grünert (2011), Morel et al. (2012), Morel (2016), Frick (in print), Stark (2012), Robert-Tissot (2015), Ueberwasser (2013).[4]

In contrast to this, and quite surprisingly given the omnipresence of mobile messengers in our daily lives, systematic research on WhatsApp messages is, at the time of writing, quite lacking. Some of the relatively early examples include Schnitzer (2012, PhD-Thesis, Munich) with a short chapter about spelling features in German WhatsApp messages compared to text messages (emoticons, punctuation, spelling mistakes, automatic error correction); or Law (2012), a small corpus-based study on mistakes among Chinese English Learners in a WhatsApp group chat with an English teacher (data-driven learning); in fact many recent studies put forward WhatsApp as a learning context in second language learning (see e. g. Hafner/Li/Miller 2015). Outside linguistics, Blasinski (2013) presents a sociological study on social interactions in romantic relationships in WhatsApp (German), or Church/de Oliveira (Telefonica Research, 2013) a sociological analysis of the perceptions of usage and motivations of users of WhatsApp and text messages (Spanish). Calero Vaquera (2014) compares

---

[2] We wish to thank Franziska Stuntebeck, Karina Frick, and Joan Miralles for bibliographical support for this sub-chapter.

[3] Terminology is also an issue; we use the traditional terminology in what follows, i. e. CMC = computer-mediated communication (Baron 1984, Herring/Stein/Virtanen 2013, Thurlow/Poff 2013); but would like to also mention CMD = computer-mediated discourse (Herring/Androutsopoulos 2015) or DD = digital discourse (Thurlow/Mroczek 2011), as well as EMC = electronically mediated communication (Baron 2008, Panckhurst/Marsh 2011) and the very useful term KSC = keyboard-to-screen communication, which explicitly includes mobile communication and excludes communication input via audio and video technologies (Jucker/Dürscheid 2012).

[4] For a complete list of our publications see: www.sms4science.ch.

text messages, WhatsApp messages and *MSN* (originally *The Microsoft Network*, today a collection of Internet services and apps for Windows and mobile devices provided by Microsoft) on the extratextual, paratextual, and intertextual level according to the use of emoticons and emojis and interpersonal relations (Spanish). Arens (2014, BA-Thesis, Münster) describes multimedial elements (audios, videos, pictures, pictograms, hyperlinks) in WhatsApp chats, again in comparison with text messages (German). More recently, some serious studies on WhatsApp have been conducted, such as the systematic comparison of text messages and WhatsApp messages (focus on German) by Dürscheid/Frick (2014), one chapter on the use of emojis in (Swiss) German WhatsApp chats in Frick (2015), or the comparison between different non-standard spelling strategies in English text messages and WhatsApp chats by Tagg (2016); spelling issues are also a topic in Sánchez Martínez (2015) or Vazquez-Cano/Mengual-Andres/Roig-Vila (2015). Apart from sociolinguistic, interactional or discourse analytical studies such as Pérez-Sabater/Montero-Fleta (2015), Sánchez-Moya/Cruz-Moya (2015) on discursive practices in Spanish WhatsApp status notes and König (2015) on dialogue structuring in German text messages and WhatsApp messages, only Meier (2015, MA-Thesis, Zurich) and Imo (2017) have worked on linguistic phenomenena in WhatsApp messages in a narrow sense; Meier (2015) more precisely on argument drop in French and German WhatsApp messages, in order to find out whether their use and distribution are triggered by technical, communicative or grammatical factors (cf. Meier/Stark accepted).

The above discussion clearly highlights that comparing text messages and WhatsApp messages is a major issue, and the two devices of mobile graphic communication in fact differ considerably, as has been shown convincingly in Dürscheid/Frick (2014). The most important technical features of WhatsApp are the internet-based technology, the unlimited number of characters available, free attachment of multimodal elements (pictures, sound files, videos), and the virtual keyboard, comprising a large selection of emojis and facilitating typing, in contrast to the old mobile phones with their 10 digit keyboard and multifunctional keys. As regards communicative affordances and expectations, WhatsApp chats resemble face-to-face-chats much more than older exchanges via text messages, as can bee seen from Table 1:

Asynchronous (CM) exchange	Quasi-synchronous WhatsApp chat
No immediate reaction expected	Temporal co-presence, direct reaction possible and also expected
Two partners, no indication whether they are online or not	Usually, all (most) partners are online (status visible)
No time pressure, no space constraints	Time pressure: quick interaction possible (and expected)
	High degree of interaction, dialogues
longer texts, near-standard, more coherent?	shorter texts, incoherent, typos, norm deviations?

Table 1: Comparison of text messages and WhatsApp chats (following Dürscheid/Frick 2014: 19–21)

The greater time pressure and expected immediate reaction in particular might lead to shorter messages, more typos, more emojis, generally more features of economy such as incomplete sentences (argument drop) – hypotheses to be investigated empirically on the basis of our WhatsApp corpus (see Section 4), which will be presented in detail in the following section.

3 The corpus

3.1 Data collection and data processing

In June and July 2014, the Swiss population was informed by several Swiss news portals (cf. www.whatsup-switzerland.ch) about the start of a public WhatsApp collection and was invited to send their chats in as attachments via an internet link and e-mail address.[5] By the end of the collection, we had received 967 chats (1,291,022 messages) and the process of cleaning them up and organizing them into a linguistic corpus was begun. As with the preceding research project on text messages, particular attention was given to protecting the senders' privacy and to integrating messages only with full consent from their respective authors. From the researcher's point of view, having all the messages within a chat available for research is imperative to pursue any kind of communicational or interactional studies, a factor that differentiates research into WhatsApp from e. g. research into traditional text messages. Thus, we provided a tool that permitted all participants to be asked to give their consent for their texts to be used for research. If they consented, they were forwarded to a questionnaire with questions on all kinds of demographic information (see below, section 3.3.4). Messages from people who only consented but did not fill in the questionnaire are available in the corpus, but without demographic data. If people did not consent, we replaced their parts in the communication with e. g. *redactedQ51tokens248characters*. This allows the prospective researcher to recognize that the message was considerably long (51 tokens, consisting of 248 characters) without any information about the content of the message being communicated.

As a next step, we anonymized the messages by replacing personal information. We followed the methods used in the establishment of the Swiss corpus of text messages (cf. Stark/Ueberwasser/Ruef (2009–2015) for the corpus and Dürscheid/Stark (2011: 309) or Ueberwasser (2015) for the anonymization)). First names in the text were rotated, i. e. where *Peter* appears in the original texts, we replaced that word with *Paul*, for example, while *Suzanne* was replaced with another female name like *Maggie*, etc. This approach allows us to keep the texts readable. Furthermore, because the same name was always replaced by the same name, researchers can recognize situations where the communication partners talk about the same person again without giving any hint as to who this person is. We did attempt to consider gender when rotating names, however this is difficult in a multilingual corpus because some names differ in gender (e. g. *Andrea* being a female name in German but a male one in Italian and Romansh). Other personal information such as last names and addresses where replaced by place holders (*[LastName]* or *[Address]*). Numbers were replaced with *N* if longer than two digits, so a general Swiss phone number would look like *NNN NNN 12 92*, as the long groups of digits were replaced but not the short ones. If somebody wrote that they are 100% sure of something, the data in the corpus will state that she is NNN% sure. Once the desired level of confidentiality was achieved, we had to clean up the data, removing duplicated chats, sent in twice by two different chat partners, or chats without linguistic content, only containing attachments (pictures or videos, not included in the corpus).[6]

---

[5] Our thanks go to Cédric Krummes and Charlotte Meisner, who helped greatly with the organization of our data collection, and to Rowan Gough, who handled the technical part of the data collection.

[6] Additionally one chat in a Southern Slavic language was removed. The chat is not long enough to be interesting for linguistic research.

WhatsApp messages are renowned for their use of emojis. In the browser, in which we make the corpus available, these graphical symbols can be represented and also individually queried. It is possible, for example, to search all messages containing a 😢 . However, if research on emojis is to be performed, it is also desirable to query for all crying faces, for example. To that end, we introduced descriptions as they are used by the Unicode agency. The emoji shown above can thus be found with a query *emojiQcryingFace*, but, using RegEx[7], also with *emojiQcrying.** The latter query will also find the crying cat face. In this way, researchers interested in emojis can query all emojis or groups of similar emojis, etc.

The example of *mediaQremoved* and *emojiQcryingFace* already introduces the concept of using category plus the uppercase letter *<Q>* plus a description to annotate non-linguistic messages. For the sake of completeness, it must be clarified that there are more of these messages, e. g. *actionQuserIN* for messages similar to "XY joined the chat". These are messages not written by participants, but generated by WhatsApp. A similar and self-explanatory approach was taken for other actions such as changes to the users' avatar, etc. We left all messages without text (i. e. action messages, encoded messages without permission, messages that contain only media and some encrypted messages, 461,919 in total) in the corpus such as to give a complete idea of the conversation including when a participant comments on a new avatar, for example. However, in the figures presented below, these are not taken into consideration.

The corpus is multilingual and the research will also be multilingual (as in the preceding project on Swiss text messages). Therefore, it is important to know the languages contained in a chat. An automated annotation of languages based on tools for natural language processing is out of the question because many messages are too short and thus do not contain enough information for a trained tool to recognize a specific language. Moreover, there is too much code switching and too much non-standard spelling at the current time. We therefore decided to use student assistants to help. These assistants looked at the first 100–500 messages of each chat to define the languages that can be found in the chat, with a main language (=100 or more messages in a specific chat) and other languages found. Languages annotated are: Swiss German dialect (GSW), non-dialectal German (DEU), French (FRA), Italian (ITA), any variety of Romansh (ROH), English (ENG), Spanish (SPA) and any Slavic language (SLA).

As a next step, the messages were tokenized, i. e. individual tokens (~words) were marked as independent units (= any expression between two spaces or after a punctuation sign, common in automated tokenization, with special rules applying by language e. g. for apostrophes). In our corpus, because of the multilingual nature and because of the unconventional spelling, we had to adjust the tokenization such as to keep emoticons together as tokens. Other adjustments to a standard tokenization process relate to money (*48.-*), time stamps (*22:30h*), apostrophes (*c'est* in French or *dell'* in Italian), hyphens (*a-t-il* in French), etc. The tokens created in this way can be queried as individual units in ANNIS.

---

[7] A syntax often used in programming to define strings for flexible pattern matching and searching.

3.2   Browsing the corpus with ANNIS

The WhatsApp-corpus in its final form will be rather complex because there will be several layers of annotation (essentially part-of-speech and a normalized layer in a standardized language plus the original messages) and because of the messages being embedded in chats. Most corpus browsers available present the data in the corpus in a linear way with individual units being marked as annotations. The most common representation in ANNIS[8] is linear, too, as presented by 1) in Figure 1. However, this type of presentation, is not useful when browsing through messages in context. Even though the author is marked above the texts, a vertical representation similar to what the chat would have looked like in WhatsApp is much more favorable. To the best of our knowledge, ANNIS is the only browsing tool that allows for the presentation to be adjusted in such a way. This vertical representation covers the whole chat, if needed. One chat can thus be browsed from beginning to end.

Apart from these special needs regarding visualization for our text type, ANNIS offers AN-NIS Query Language, a search syntax that allows users to search across layers and for complex structures. By using ANNIS as a browser, we have a very powerful query-language available, we can display the chats in a very efficient way and we will be able to make the corpus available on the Internet.

Figure 1: Example of messages presented in ANNIS. 1 showing a linear representation, with messages from different authors following each other, while 2 represents the chat similarly to how it would be seen by the informants

---

3.3   Facts and figures

As noted above, the corpus is still being worked on. The facts and figures presented below represent the corpus as it stands in April 2017 (Release v4.0). At this point, the corpus is already very mature, yet minor changes may still be made. Once the corpus is available to a general public, an integrated documentation will reflect the latest figures.

3.3.1   Basics

The corpus consists of 1,188,570 messages containing written text in 617 chats that were written by approximately 1,538 participants (see Section 3.3.4 to understand why this figure is only an approximation). 945 participants gave their consent to use their 763,650 messages consisting of 5,543,692 tokens. A total of 426 participants provided additional information about their demographics in the questionnaire. Every chat can be a form of communication between two or more people. One chat in our collection was written by 31 participants, but unfortunately we do not have the consent of all participants to use the messages in this chat.

In Table 2, we arranged these figures by language, taking only Swiss national languages into consideration. As mentioned above, we differentiate between languages that can be found in more than 100 messages per chat and those that are less frequent. In Table 2 (and further presentations) when quoting figures per language we look at the total length of each individual chat. If a chat is shorter than 100 messages, we consider all languages that can be found in this specific chat and count the chat for all these languages. If the overall chat is longer than 100 messages, we only consider languages that are relevant for more than 100 messages, and again, the chat will be counted for all these languages. As a consequence, the total number of chats, messages tokens and participants quoted in Table 2 is higher than the actual figures available in the corpus. Additionally, as only those messages, for which we have the participants' consent will be used for research, we will disregard messages without consent for all following figures.

	DEU	GSW	FRA	ITA	ROH
Messages with consent					
Chats	93	275	141	87	77
Messages	81,456	506,984	197,255	42,559	29,094
Tokens	625,419	3,611,033	1,397,375	293,567	283,909
Participants	159	444	188	133	122
Chat with most participants	6	10	6	8	7
Consent from everybody in the chat					
Chats	46	112	42	25	28
Messages	63,923	361,339	110,255	16,669	21,379
Tokens	453,911	2,670,685	756,993	126,569	191,805
Participants	106	259	84	60	66
Chat with most participants	6	9	2	7	7
Demographics from everybody in the chat					
Chats	25	59	32	11	13
Messages	24,426	213,598	103,452	5,268	7,271
Tokens	177,220	1,822,271	705,061	39,649	80,788
Participants	50	125	64	25	26
Chat with most participants	2	9	2	5	2

Table 2: Number of chats, messages and participants for all chats, chats with consent from all participants, demographics from all participants

The first part of Table 2 shows all messages for which we have the consent of participants, i. e. messages that can be used for research. The second part presents the data relevant for discourse analytical research since it shows only chats, for which we have the consent of all participants of a specific chat. The third part presents the data that is of interest for sociolinguistic studies, for example. All participants in this part provided demographic data in addition to their chats, for which we of course also obtained consent.

In all three parts presented in Table 2, Swiss German Dialects provide most data, while the data for Italian and Romansh are not so abundant, though should still be sufficient for interesting research.

3.3.2 Length

The individual chats vary greatly in length. On the one hand, there are 277 chats that are shorter than 100 messages. On the other hand, the longest chat is 29,238 messages long (i. e. we have the consent for that number of messages). Long chats, however, are the exception, as can be seen in Figure 2. The average length of chats differs by language, with the average chat in a Swiss German dialect being more than four times longer as the average one in Romansh. Chats in a Swiss German dialect are therefore not only more frequent than chats in any other language or variety, they are also longer in our data.
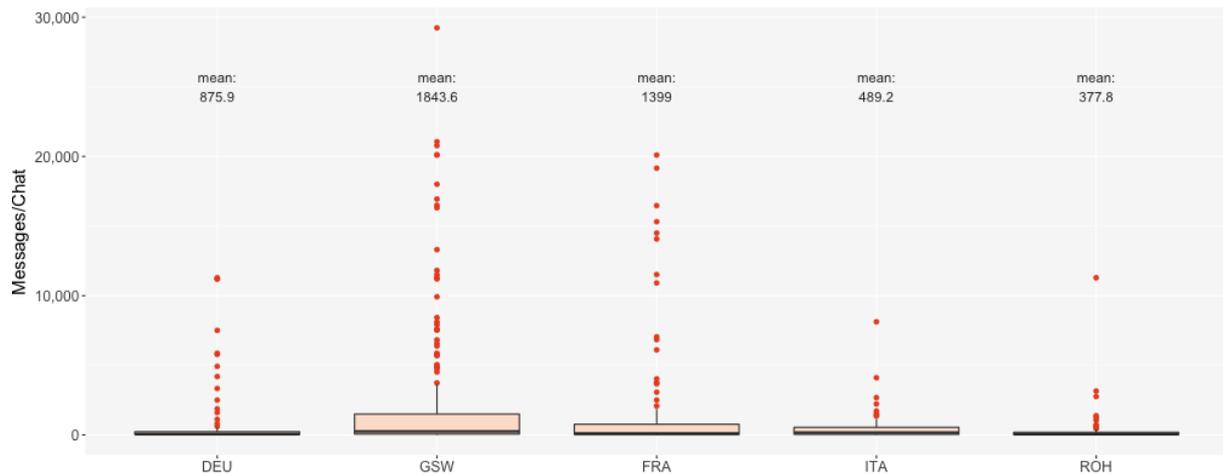
Figure 2: Messages per chat per language

The length of the individual messages is much more leveled between the languages, as illustrated in Figure 3. For all languages, the average length of the messages is around thirty characters. As a comparison, in the Swiss corpus of text messages, this figure is 115 characters. WhatsApp messages are thus considerably shorter than text messages in a comparable corpus. Even though the messages in the Swiss German part of the corpus are on the shorter side compared to Romansh or non-dialectal German, the single longest message of 10,487 characters can be found in this part of the corpus. This message consists of one single exclamation of *AAA* interrupted by some <*À*> and spaces.
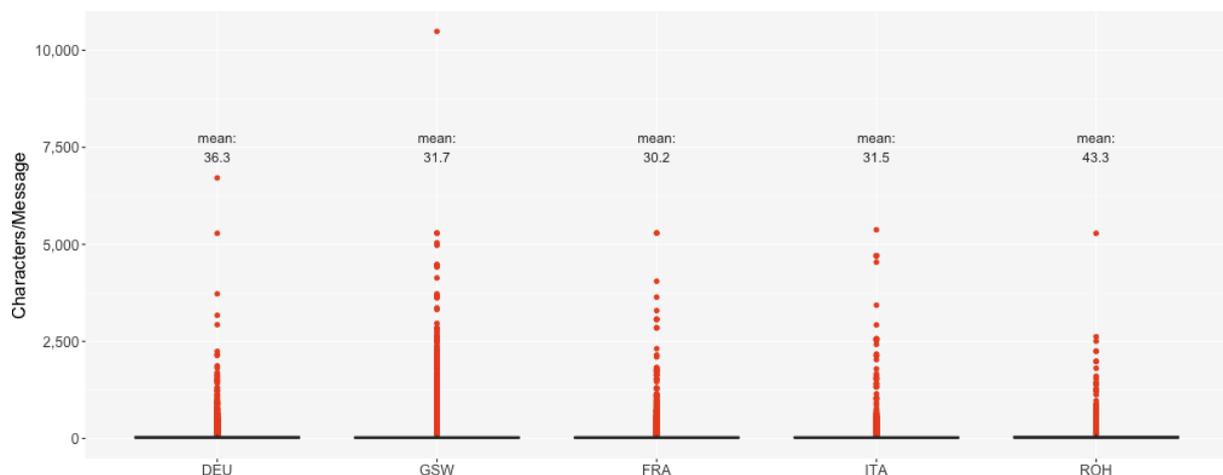
Figure 3: Characters per message

Overall, long messages are the exception. 25% of all messages are shorter than 8 characters, 50% shorter than 20 characters and 75% shorter than 38 characters. Figure 4 shows only messages shorter than 20 characters. As can be seen, French, Swiss German and Romansh are strong in messages with only one character, often emojis. Generally, all languages in the corpus show a very similar distribution in this respect, even though the tendency towards extremely short texts is stronger in dialectal German and in French compared to the other languages/varieties. The fact that we have more messages from these languages may influence this figure, of course.

Considering Figure 4, a peak in messages with 6 characters, especially for GSW and ITA, can be seen. This extensive use of messages with 6 characters is caused by forms of *hahaha* either

in upper or in lower case. This exclamation accounts for 13.3% of all occurrences of six character messages in all languages.
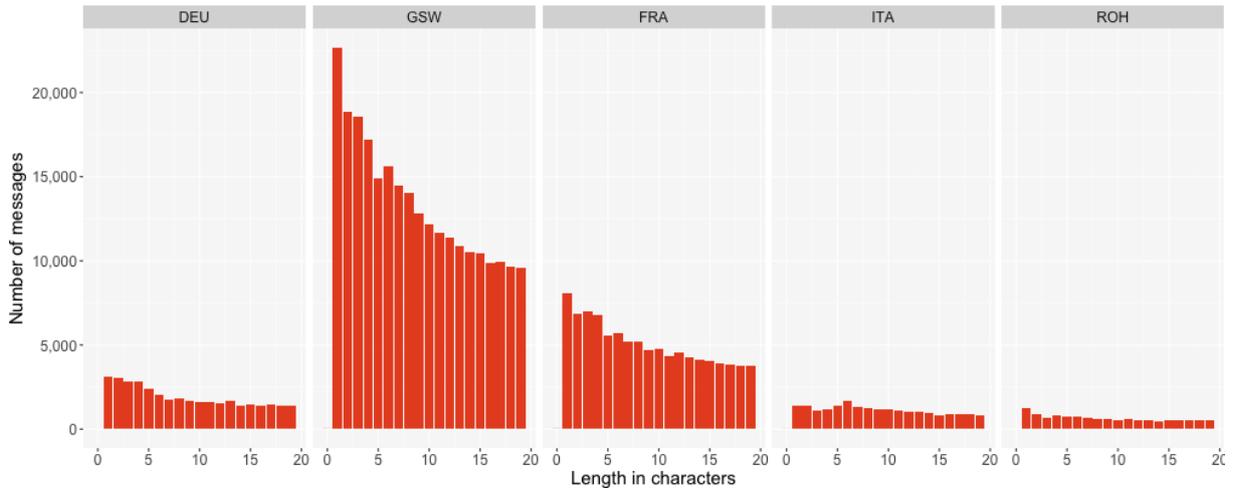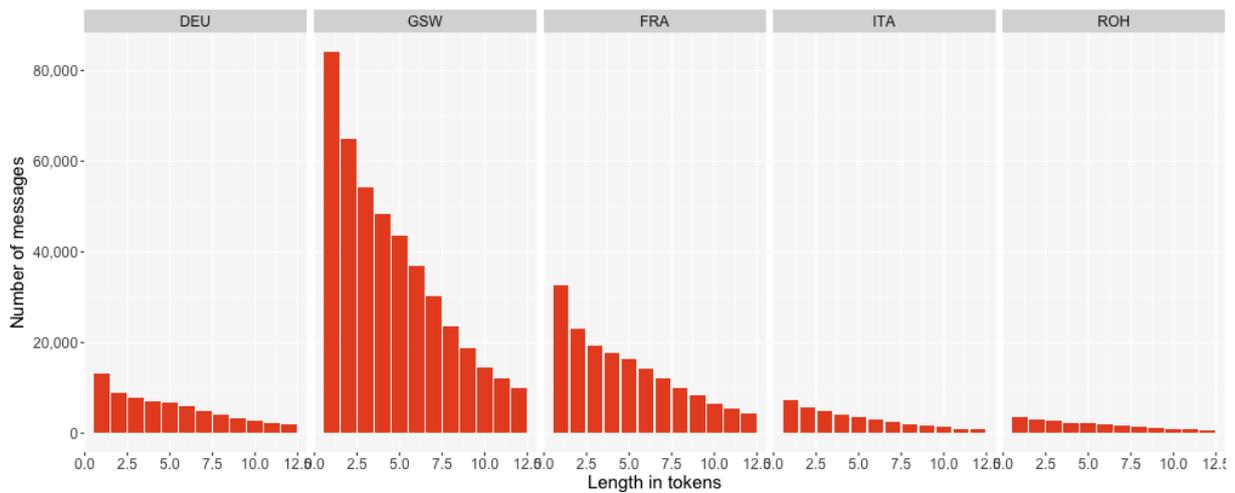

Figure 4: Characters per message shorter than 20 characters

Citing the size of a corpus in characters is not common – the normal dimension would be tokens. However, tokenizing a corpus like ours is not all that easy as can be seen from the message with the most tokens in the corpus. This message is in Italian and starts with the sentence *Scusate, passa un millepiedi* ('sorry, there is a millipede passing by') and then 344 lines similar to *..??(???)??*. Together, these lines form a zigzag line. The tokenizer counts this messages as 3,358 tokens long, a number that is as good as any other, since nobody can really recognize tokens in this message. This *millipede* accounts for six out of the 10 messages that are longer than 1,500 tokens (with the introduction in Italian, in both varieties of German and without any text). The other four messages longer than 1,500 tokens are repetitions of crying or loving emojis.
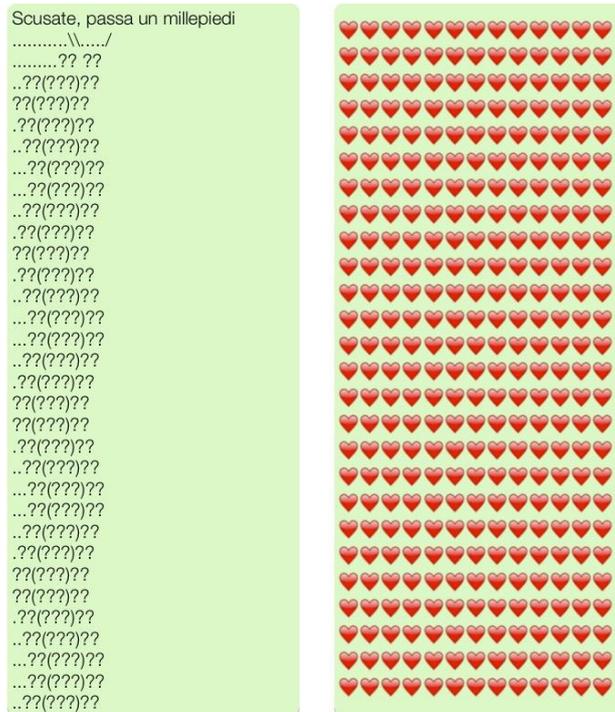

Figure 2: Examples of long chats (shortened): Millipede and hearts

are, the French and non-dialectal German data feature the highest number of emojis for the group between 50 and 64 years of age. Is there an isolated participant in this group who is keen on emojis or do other factors influence this distribution? We must leave this question open for the moment (but see sub-project B described in Section 4).
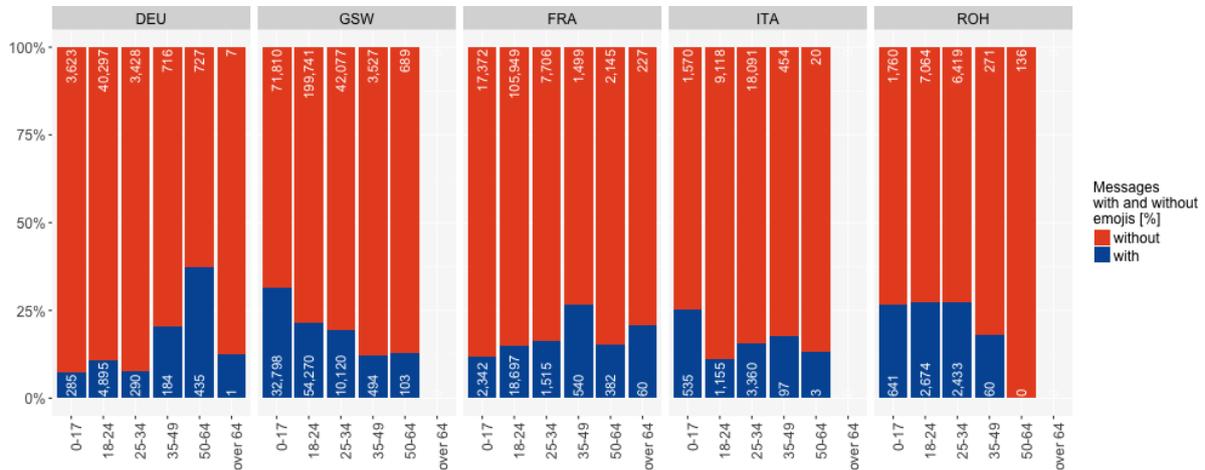
Figure 7: Percentage of messages with emojis by age and language

Figure 8 can perhaps offer some insight into this question since it shows the use (or lack thereof) of emojis for every participant with demographic information in our data set. As expected, the use of emojis in this chart declines with age, meaning it must be assumed that a single individual using many emojis in their text is responsible for the high number of emojis for their age group in Figure 7.
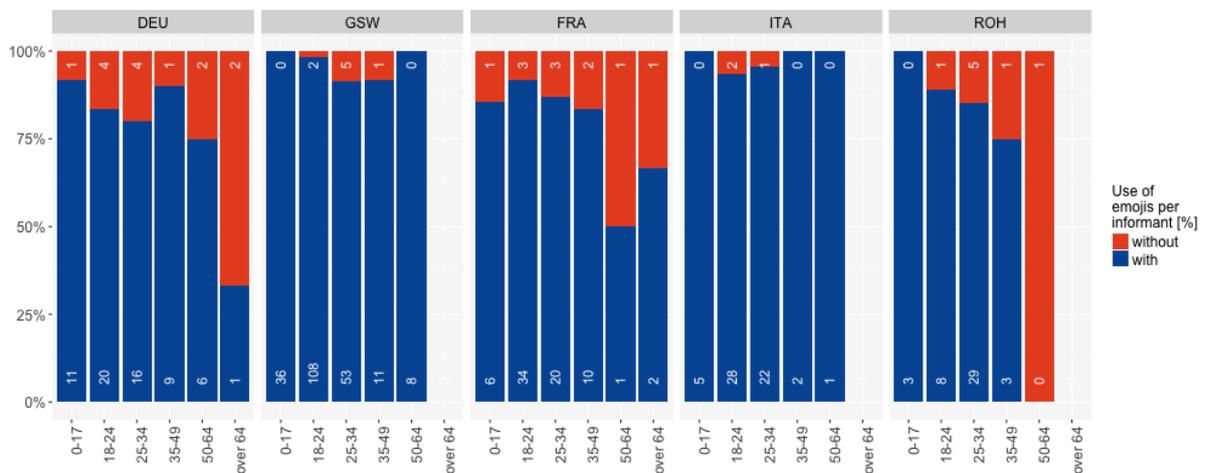
Figure 8: Percentage of participants using emojis by age and language

3.3.4   Demographics

426 participants filled in an online questionnaire containing questions regarding age, gender, education, profession, input method (i. e. prediction and auto correction), mother tongue and language use, place of residence in 5th grade, post code of work and residence. The data they provided not only allow us to know more about our participants, but also to know how many chats they are involved in because we can identify the person even if they are only a chat partner and did not send in the chat themselves. The general procedure when submitting chats

was such that one person submitted the chat and was then requested to a) fill in the demographics questionnaire and b) inform his/her chat partners to give their consent and fill in the demographics questionnaire as well. However, depending on the reaction of the chat partner(s), we are not allowed to use their contributions or even have their demographics.

3.3.4.1 Age

The participants who sent in messages and answered the questionnaire are mostly younger than 35, as can be seen in Figure 9. For Swiss German dialects and non-dialectal German, as well as for French, the group between 18 and 24 is not only the strongest, but also the most productive one. The very young and rather old groups, on the other hand, are unfortunately not particularly present. The collection was advertised in the press but also at the participating universities. As a consequence, many participants are in fact students.[9]
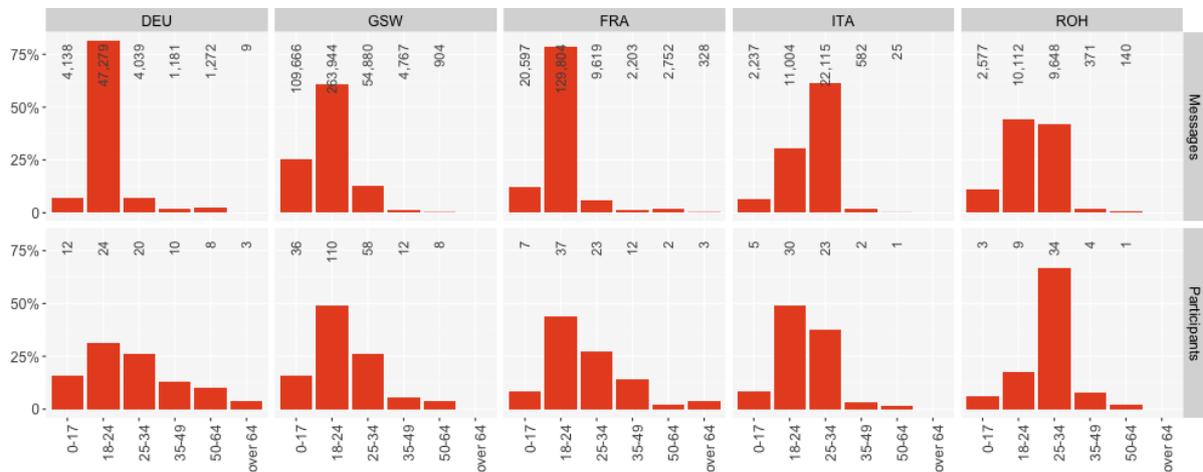
Figure 9: Age of participants per message and per participant

3.3.4.2 Gender

Figure 10 splits the data from Figure 9 further by integrating gender. As we can see, the number of male and female participants is rather balanced over the whole data set, and the same can be said for the number of messages submitted. This is not the case, however, for age groups where only a few people filled in the questionnaire, especially for the older groups. Since we have only one participant each for Italian and Romansh in the available oldest group, their respective gender dominates this age group. For standard German there are three participants in the oldest group, but since these participants only submitted 9 messages, they do not have a strong influence on the data as a whole. While female participants are slightly dominant in both varieties of German, they clearly contributed fewer chats than males in the dialect as can be seen in the second chart in the top row. A similar picture can be offered for Italian messages. Here, too, we see a rather balanced situation between participants but with males contributing more chats. In the French data, on the other hand, male participants are

---

[9] Participating students also advertised the project among their peers, as can be seen from the following message (626,755) in the corpus: *Wärsch iiverstande mitzmache? Es isch äs projekt vo mire uni wo whatsapp nachrichte untersuecht :p* ('Would you agree to participate? It is this project run at my university investigating WhatsApp messages :p').

dominant in the younger and in the very oldest age range, while female participants dominate the group between 35 and 64.
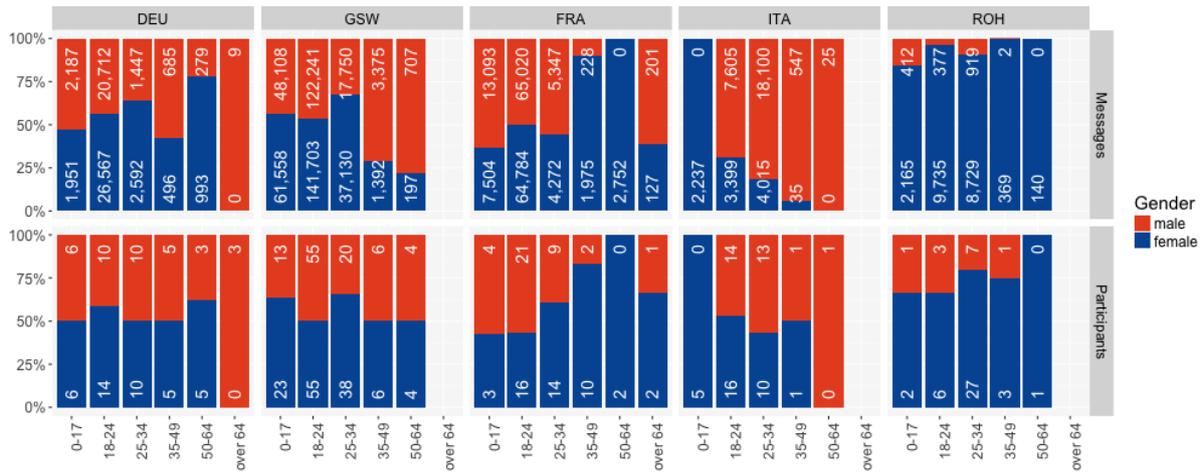
Figure 10: Gender of participants by age group and per message and participant

3.3.4.3 Education

In general it can be stated that most people participating in our data collection have a university diploma, but that those still in education submitted the most messages. Since the project was highly advertised at the participating universities, it must be assumed that it was many members of these organizations who submitted their chats, a fact that is supported by Figure 11.
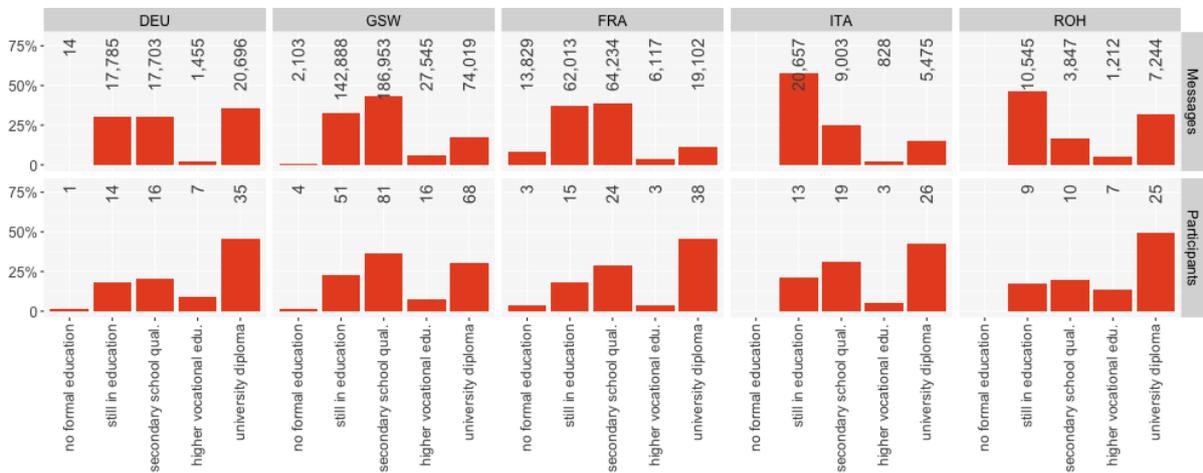
Figure 11: Education of participants by age group and per message and participant

3.3.4.4 Regional origin of messages

For most participants who provided demographics we know where they live, where they work and where they lived when they were in fifth grade, i. e. about 11 years old, since they provided the according postal codes. We then assigned these postal codes to the corresponding municipalities. This relationship is not always one-to-one. Big cities, like Zurich, are divided up into more than 30 postal codes. Some postal codes, on the other hand, belong to more than one municipality. The postal code 8127, for example, belongs to Maur and Küsnacht. The

third complication is represented by consolidated municipalities. The originally more than 20 municipalities of the canton of Glarus were consolidated into only three municipalities in 2011, yet they kept their postal codes.
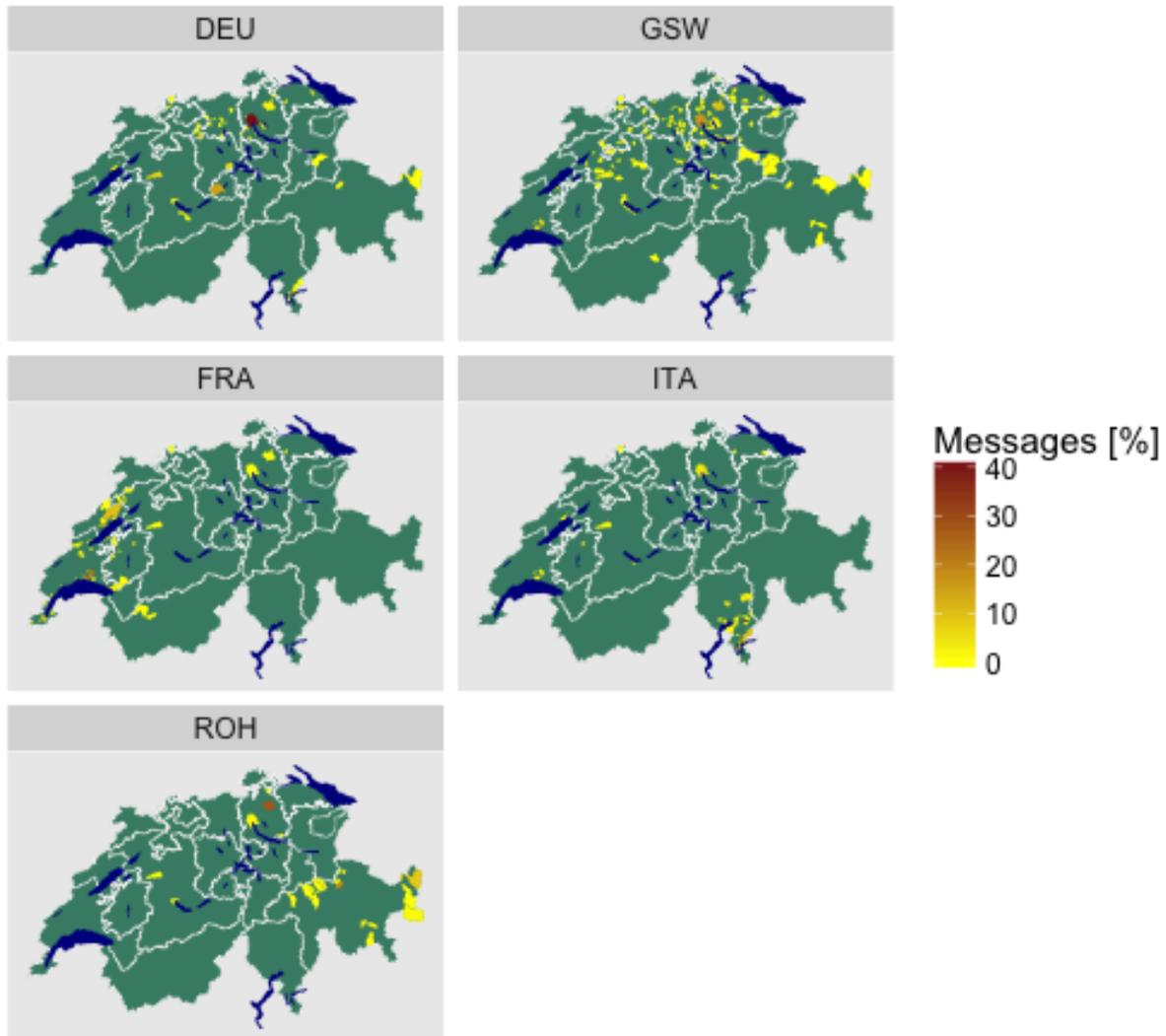
Figure 12: Number of messages per home municipality of participants

Figure 12 shows the number of messages in each language that were written by people living in a specific municipality. It comes as no surprise that the city of Zurich is strong in providing messages in all languages/varieties. When comparing standard to dialectal German, Zurich is much more dominant in providing standard German messages than dialectal ones. French data, on the other hand, only come from French speaking parts as well as from three German-speaking cities: Zurich, Basel and Winterthur. Data in Italian are more spread out among the language regions, coming from the German and the French-speaking parts of Switzerland in addition to the Tessin. Since the majority of these data come from cities with universities (Zurich, Bern, Basel, Neuchâtel and Lausanne), we can assume that a large part of the Italian data was provided by students. The biggest surprise is the data for Romansh which comes not

only from the Romansh-speaking area, but also from Bern, Thun and Zurich. By far most messages were submitted by people living outside the Romansh area in Winterthur.

Internal migration is the focus of Figure 13 which shows migration paths comparing the participants' place of residence in fifth grade to their current place of residence. Clearly, Zurich and Bern are the areas where most people move to, while our participants tend to move away from and not towards the other major German-speaking city, Basel. In the French-speaking part, Lausanne, not Geneva, is the place to go, even for participants who grew up in the Italian-speaking part. However, many more participants from that part of the country moved to Zurich. Chur looks like a node, too, but here, more than in the other nodes, traffic goes towards as well as away from the city. It appears as if people from the Romansh-speaking valleys move to Chur, while people who grew up there move away to the Zurich area.
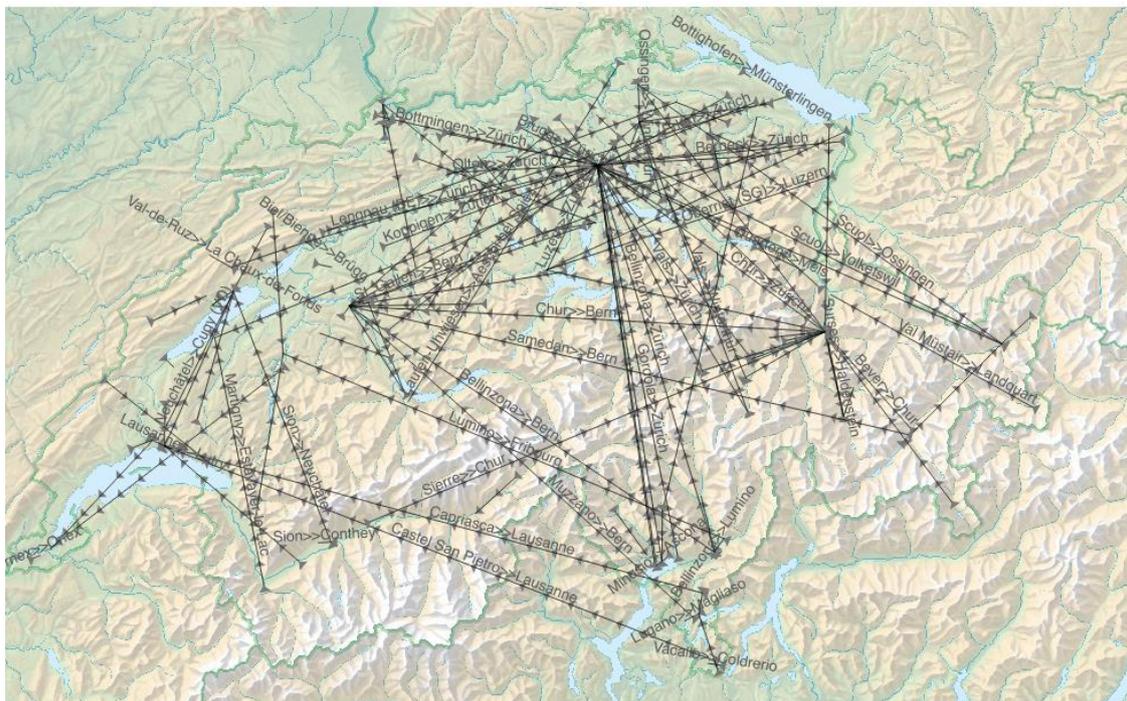
Figure 13: Internal migration between 5th grade and today

In addition to this internal migration, we have 13 participants who spent their 5th grade in Germany, 6 in Italy, 3 in France, 2 in the Czech Republic and in Croatia each, and one each in Belgium, Canada, Finland, Turkey and Luxembourg.

3.3.4.5 Typing assistance

Contrary to modern smartphones, old mobile phones used a keyboard with ten digits, but did not offer the option to install applications (apps) like WhatsApp (see Dürscheid/Frick 2014). We can therefore assume that all participants used smartphones to type their messages. The layout of the smartphones' keyboard does not have to be the same between brands, and because of the option of installing keyboards in different languages and even third party keyboards like *SwiftKey* we have no way of knowing exactly how our participants typed their texts. However, the layout of the keyboard is not the most important factor influencing the typing style. Much more important is the use of the correction tool (i. e. the smartphone adjusts typing errors, correcting to *I've* when we type *Ive*, for example) as well as prediction

(e. g. suggesting *home*, *homework* and *homage* when users type *hom*). Availability and usefulness of either function depend on the operating system, the keyboard software, and the language used.

Table 4 shows the information we received from our participants concerning the use of correction and prediction. People who stated that they do not use these features and those that did not provide any answers for these questions are pooled together to calculate the percentages, assuming that they would have ticked *yes* if they used the respective feature.

	DEU	GSW	FRA	ITA	ROH
Messages: correction	7,039 (12.2%)	15,208 (3.5%)	61,946 (37.5%)	25,593 (71.2%)	71 (0.3%)
Messages: prediction	2,434 (4.2%)	30,527 (7%)	34,762 (21%)	17,083 (47.5%)	4,470 (19.6%)
Informants: correction	14 (18.2%)	15 (6.7%)	49 (58.3%)	31 (50.8%)	2 (3.9%)
Informants: prediction	12 (15.6%)	15 (6.7%)	24 (28.6%)	17 (27.9%)	3 (5.9%)

Table 4: Participants' use of correction and prediction per message and per participant

It comes as no surprise that the use of both features is much more frequent for participants writing in French and Italian, because there is no support for the Swiss German dialects nor for the different Romansh varieties. The use of these features is higher in messages with German non-dialectal content than in those with dialectal messages for the same reason. The fact that the values for non-dialectal German are still much lower than for French is most likely due to code-switching between the dialectal and the non-dialectal variety. The only outstanding figure is that for messages written with prediction in Romansh. Out of the 4,470 messages with this feature, 4,397 are written by one and the same participant. Most of these messages (3,078 messages) are found in chats that are marked as bilingual (GSW, ROH) and are of an informal character, which may mean that the prediction is switched on for standard German text (parts).

3.4   Further steps

The corpus as it is described here is available for our own project-internal research at the time of writing and – upon request – also to outsiders for academic purposes. Work on the corpus continues, of course. As a next step, we wish to add a normalized layer, i. e. an annotation on a per word basis that gives the value of the word in the respective standardized spelling. Alongside this task, we are working on a part-of-speech annotation. The two tasks are of course interdependent. Having a part-of-speech annotation available will improve the correct alignment of non-standard homographs to the standardized spelling, while knowing the correct spelling will help with the annotation of the correct part of speech. Neither of these jobs can be accomplished by humans because of the size of the corpus (5,543,692 tokens). We have therefore decided to choose a mixed approach and will use student helpers to annotate some of the data and then use trained tools from computational linguistics to annotate the rest of the corpus.

By the end of the research project "What's up, Switzerland?" (end of 2018), the corpus will be made fully accessible online.

4   Investigating the corpus

We saw in Section 2 that systematic research on WhatsApp messages and large-scale corpora of WhatsApp data have so far been lacking. This is regrettable for at least two reasons: it has become an omnipresent form of mobile messaging, see also the intense media coverage of the topic (Google hits as of 23/10/2016: *Neue Zürcher Zeitung*: 5,410; *Le Matin*: 2,330; *Tagesanzeiger:* 32,300), and it systematically includes an intense interactive dimension (for its relevance, also for linguistic change, see Herring 2012, Trudgill 2014). Facing this lack of knowledge about the nature of WhatsApp communication, the SNSF-funded Sinergia project "What's up, Switzerland?" has two research objects; the messages themselves and their linguistic make-up on the one hand; and the discourse on WhatsApp messaging, on the other. Besides quantitative approaches to different linguistic phenomena, qualitative analyses of *individual statements* in speaking and writing about WhatsApp (users and the media) are also analyzed. Two main research questions guide our research work (cf. the project internet site: www.whatsup-switzerland.ch): 1. What do Swiss WhatsApp messages look like? What has changed overall between Swiss SMS and Swiss WhatsApp messages, and why (as regards linguistic structures, use of images in a broad sense, spelling, register-specific style, individualization vs. accomodation)? 2. What is said / done by the individual users and the media in/on WhatsApp messages and chats, in relation to the findings for question 1? The project involves researchers from four universities, under the direction of Elisabeth Stark (University of Zurich): Bruno Moretti/Silvia Natale (University of Bern), Christa Dürscheid (University of Zurich), Federica Diémoz (University of Neuchâtel), Beat Siebenhaar (University of Leipzig, Germany), and Crispin Thurlow (University of Bern).

Four sub-projects will investigate different aspects of Swiss WhatsApp communication in the four national languages of Switzerland and their varieties. *Sub-project A, "Language(s) of WhatsApp: Verbal Periphrases and Argument Drop"*, directed by Elisabeth Stark (Zurich) and Silvia Natale (Bern), with two doctoral students (Franziska Stuntebeck, Zurich, Rosella Maraffino, Bern) continues work partially undertaken in the preceding SMS-project by focusing again on the morphosyntax of the messages under investigation, a domain strongly neglected by international research in CMC. More precisely, we will analyze patterns of argument drop (already investigated in the SMS project, see Frick 2015; Robert-Tissot 2015), and the use of progressive verbal periphrases, in a thoroughly cross-linguistic perspective. We intend not only to describe but also to explain the patterns we will find in order to understand whether these are register-specific features (in the sense of Biber 1995) or mainly external, i.e. technologically provoked structures. We are also interested in the role context plays and the limits of variation (see Meier/Stark (accepted) for a pilot study on argument drop in French and German WhatsApp messages). *Sub-project B, "Language Design in WhatsApp: Icono/Graphy"*, directed by Christa Dürscheid (Zurich) and Federica Diémoz (Neuchâtel), with two post-docs (Christina Siever, Zurich; Etienne Morel, Neuchâtel) looks at graphematic issues in a broad sense. Continuing work done by Marie-José Béguelin in the SMS-project (see Béguelin 2012), the Neuchâtel team will analyse different spelling strategies in French and German WhatsApp messages, whereas Christa Dürscheid and her post-doc will systemat-

ically analyze the distribution and function of emojis in our data (see section 3.3.3). *Sub-project C, "Individuals in WhatsApp",* directed by Beat Siebenhaar (Leipzig), with doctoral student Samuel Felder, will take a look at individuals in chats and their way of accommodating or rather distinguishing themselves from their chat partners with repect to the linguistic and graphematic variables investigated in sub-projects A and B, plus code-switching patterns, thoroughly described in the SMS-project (see Cathomas 2015; Cathomas et al. 2015; Bucher 2016; Morel 2016; Morel et al. 2012). Finally, *sub-project D, "Ideologies: The Cultural Discourses and Social Meanings of Mobile Communication",* directed by Crispin Thurlow (Bern), with doctoral student Vanessa Jaroski, will describe (also by setting up a "Digital Discourse Database", http://www.crispinthurlow.net/digital-discourse-database.php) and analyze the public discourse on graphic mobile communication via WhatsApp (and SMS), attempting to pin down the way Switzerland looks at the revolutionary developments in our communicative behavior and its evaluation by the media.

In all four sub-projects, the comparative approach, i. e. the close examination of data from the four national languages of Switzerland and some of their varieties, is crucial.

Since its start in January 2016, substantial work on the corpus data (see sections 3.1 and 3.2) has been undertaken and the corpus is ready to be analyzed by the project's team at the time of writing. Temporary access from outside can be requested for scientific purposes only by contacting the authors. Initial results were presented at the opening workshop in June 2016 at the University of Zurich (http://www.whatsup-switzerland.ch/index.php/en/research-en/workshop1), and the upcoming events, seminars, student papers, talks and publications on the project are available on the project's website (http://www.whatsup-switzerland.ch/index.php/en/research-en/talks-en).

We are confident that over the next two and a half years we will unravel basic linguistic, graphical, variational and discursive properties of WhatsApp messages in Switzerland and more generally, in this first large-scale research project on one of the most important forms of mobile communication of our times.

References

Arens, Katja (2014): „WhatsApp: Kommunikation 2.0. Eine qualitative Betrachtung der multimedialen Möglichkeiten". In: König, Katharina/Bahlo, Nils (eds.): *SMS, WhatsApp & Co*. Münster, MV-Verlag: 81–106.

Baron, Naomi (1984): "Computer-mediated communication as a force in language change". *Visible Language* 18/2: 118–141.

Baron, Naomi (2008): *Always on. Language in an Online and Mobile World*. Oxford: Oxford University Press.

Béguelin, Marie-José (2012): « La variation graphique dans le corpus suisse de SMS en français ». In: Caddéo, Sandrine/Roubaud, Marie-Noëlle/Rouquier, Magali/Sabio, Frédéric (eds.) : *Penser les langues avec Claire Blanche-Benveniste*. Aix-Marseille, Presse de l'Université de Provence: 47–63.

Biber, Douglas (1995): *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Blasinski, Jennifer (2013): *Von den Nutzungspraktiken der Anwendung ‚WhatsApp' im Rahmen romantischer Beziehungen.* München: Grin.

Bucher, Claudia (2016): *SMS-User als ‚glocal player': Formale und funktionale Eigenschaften von Codeswitching in SMS-Kommunikation.* Hannover/Leipzig: Networx. http://www.mediensprache.net/networx/networx-73.pdf [02.06.2017].

Calero Vaquera, María Luisa (2014): "El discurso del WhatsApp: Entre el *messenger* y el *SMS*". *ORALIA* 17: 85–114.

Cathomas, Claudia (2015): *Von „I dont Know!" zu „Kei problem chara!!" – Eine korpuslinguistische Untersuchung zu rätoromanischen SMS unter besonderer Berücksichtigung verschiedener Formen und Funktionen von Code-Switching.* Unpublished Doctoral Dissertation, University of Bern.

Cathomas, Claudia/Ferretti, Nicola/Bucher, Claudia/Morel, Etienne (2015): "Same same but different: Code-Switching in Schweizer SMS – ein Vergleich zwischen vier Sprachen". *Tranel* 63: 171–189.

Church, Karen/de Oliveira, Rodrigo (2013): "What's up with WhatsApp?: comparing mobile instant messaging behaviors with traditional SMS". In: Rohs, Michael/Schmidt, Albrecht/Ashbrook, Daniel/Rukzio, Enrico (eds.): *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*. Munich, ACM: 352–361.

Dürscheid, Christa/Stark, Elisabeth (2011): "*Sms4science*: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland". In: Thurlow, Crispin/Mroczek, Kristine (eds): *Digital Discourse. Language in the New Media*. New York/London, Oxford University Press: 299–320.

Dürscheid, Christa/Frick, Karina (2014): „Keyboard-to-Screen-Kommunikation gestern und heute: SMS und WhatsApp im Vergleich". In: Mathias, Alexa/Runkehl, Jens/Siever, Thorsten (eds.): *Sprachen? Vielfalt! Sprache und Kommunikation in der Gesellschaft und den Medien*. Networx 64: 149–181.

Dürscheid, Christa/Frick, Karina (2016): *Schreiben Digital. Wie das Internet unsere Alltagskommunikation verändert*. Stuttgart: Kröner.

Frick, Karina (2017): *Elliptische Strukturen in SMS. Eine korpusbasierte Untersuchung des Schweizerdeutschen*. Berlin/Boston: de Gruyter.

Grünert, Matthias (2011): „Varietäten und Sprachkontakt in rätoromanischen SMS". *Linguistik Online* 48: 83–113.

Hafner, Christoph/Li, David/Miller, Lindsay (2015): "Language Choice Among Peers in Project-Based Learning: A Hong Kong Case Study of English Language Learners' Plurilingual Practices in Out-of-Class Computer-Mediated Communication". *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes* 71/4: 441–470.

Herring, Susan (2012): "Grammar and Electronic Communication". In: Chapelle, Carol (ed.): *The Encyclopedia of Applied Linguistics*. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0466/abstract [24.10.2016].

Herring, Susan/Stein, Dieter/Virtanen, Tuija (2013): "Introduction to the pragmatics of computer-mediated communication". In: Herring, Susan/Stein, Dieter/Virtanen, Tuija (eds.): *Pragmatics of Computer-Mediated Communication*. Berlin/Boston: Mouton de Gruyter: 3–32.

Herring, Susan/Androutsopoulos, Jannis (2015): "Computer-mediated discourse 2.0". In: Tannen, Deborah/Hamilton, Heidi/Schiffrin, Deborah (eds.): *The Handbook of Discourse Analysis.* Chichester, John Wiley & Sons: 127–151.

Imo, Wolfgang (2017): „*Ob*-Sätze in der mündlichen und schriftlichen Interaktion". *Deutsche Sprache* 45/1: 1–30.

Jucker, Andreas/Dürscheid, Christa (2012): "The Linguistics of Keyboard-to-Screen Communication. A New Terminological Framework". *Linguistik online* 56/6: 39–64.

Koch, Peter/Oesterreicher, Wulf (2011): *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*. 2., aktualisierte und erweitere Auflage. Berlin/New York: de Gruyter.

König, Katharina (2015): "Dialogkonstitution und Sequenzmuster in der SMS- und WhatsApp-Kommunikation". *TRANEL* 63: 87–107.

Law, Locky (2012): *Learner Corpus Mini-research: Corpus-based Analysis Whatsapp Group Chat.* Hong Kong: The Hong Kong Polytechnic University. https://drive.google.com/file/d/0B9emHQBd5COmZW1EMEhwMUxyck0/view [24/10/2016].

Meier, Petra (2015) : « Omissions des arguments verbaux dans les messages WhatsApp (français et allemands) ». Unpublished MA thesis. University of Zurich.

Meier, Petra/Stark, Elisabeth (accepted): "Argument drop in Swiss WhatsApp messages – a pilot study on French and (Swiss) German". *Zeitschrift für französische Sprache und Literatur*.

Morel Etienne (2016): *Le bricolage plurilingue dans la communication par texto: Interprétations d'une pratique entre affiliation locale et aspiration globale*. Unpublished Doctoral Dissertation. Université de Neuchâtel.

Morel, Etienne/Bucher, Claudia/Pekarek-Doehler, Simona/Siebenhaar, Beat (2012): "SMS communication as plurilingual communication: Hybrid language use as a challenge for classical code-switching categories". *Lingvisticæ Investigationes* 35/2: 260–288.

Panckhurst, Rachel (2007): « Discours électronique médié: quelle évolution depuis une décennie? ». In: Gerbault, Jeannine (ed.): *La langue du cyberespace: de la diversite aux normes*. Paris, L'Harmattan : 121–136.

Panckhurst, Rachel/Marsh, Debra (2011): « Les frontières pédagogiques sont-elles remises en question par l'utilisation des réseaux sociaux ? L'implémentation d'objets d'apprentissage sociaux dans un espace de communication électronique médiée ». In: Liénard, Fabienne/Zlitni, Sami (eds.): *La communication électronique: enjeux de langues.* Limoges, Lambert-Lucas : 293–301.

Pérez-Sabater, Carmen/Montero-Fleta, Maria Begoña (2015): "A first glimpse at mobile instant messaging: Some sociolinguistic determining factors". *Poznan Studies in Contemporary Linguistics* 51/3: 411–431.

Robert-Tissot, Aurélia (2015): *Le sujet et son absence dans les SMS français. Une analyse basée sur le corpus sms4science suisse.* Unpublished Doctoral Dissertation, University of Zurich.

Sánchez Martínez, Sonia (2015): "La escritura de los jóvenes en los chats en el siglo XXI/How young people write in chats in the 21st century/L'écriture des jeunes d'aujourdui dans leurs chats au XXIème siècle". *Didáctica : Lengua y Literatura* 27: 183–196.

Sánchez-Moya, Alfonso/Cruz-Moya, Olga (2015): "'Hey there! I am using WhatsApp': a preliminary study of recurrent discursive realisations in a corpus of WhatsApp statuses". *Procedia – Social and Behavioral Sciences* 212: 52–60.

Schnitzer, Caroline-Victoria (2012): *Linguistische Aspekte der Kommunikation in den neueren elektronischen Medien SMS – E-Mail – Facebook*. München: Grin.

Stark, Elisabeth (2011): « La morphosyntaxe dans les SMS suisses francophones: Le marquage de l'accord sujet – verbe conjugué ». *Linguistik Online* 48: 35–48.

Stark, Elisabeth (2012): "Negation marking in French text messages". *Lingvisticæ Investigationes* 35/2: 341–366.

Stark, Elisabeth/Ueberwasser, Simone/Ruef, Beni (2009–2015): *Swiss SMS Corpus*. University of Zurich. www.sms4science.ch [24.10.2016].

Tagg, Caroline (2016): "Text messaging then and now: corpus analysis of respelling practices over time". Talk given at *PLIN Linguistic Day*, Université catholique de Louvain, Louvain-la-Neuve (12.05.2016).

Thurlow, Crispin/Mroczek, Kristine (eds.) (2011): *Digital Discourse: Language in the New Media*. New York/London: Oxford University Press.

Thurlow, Crispin/Mroczek, Kristine (2011): "Fresh Perspectives on New Media Sociolinguistics". In: Thurlow, Crispin/Mroczek, Kristine (eds.): *Digital Discourse: Language in the New Media*. New York/London, Oxford University Press: XIX–XLIV.

Thurlow, Crispin/Poff, Michele (2013): "Text messaging": In: Herring, Susan/Stein, Dieter/Virtanen, Tuija (eds.): *Pragmatics of Computer-Mediated Communication*. Berlin/Boston, Mouton de Gruyter: 163–190.

Trudgill, Peter (2014): "Diffusion, drift, and the irrelevance of media influence". *Journal of Sociolinguistics* 18/2: 214–222.

Ueberwasser, Simone (2015): *The Swiss SMS Corpus. Documentation, facts and figures*. www.sms4science.ch [24.10.2016].

Vazquez-Cano, Esteban/Mengual-Andres, Santiago/Roig-Vila, Rosabel (2015): "Lexicometric Analysis of the Specificity of Teenagers' Digital Writing in Whatsapp". *RLA, Revista De Linguistica Teorica y Aplicada* 53/1: 83–105.