Normalization

The spelling of many tokens in the SMS corpus deviate from the standard spelling rules of the according language and consequently, automated data processing such as the annotation of parts of speech (PoS) is difficult if not impossible. In order to overcome these problems, a second, normalised parallel corpus was created per language. In this normalised corpus, every token is spelled in a way that can be interpreted by tools used in computational linguistics.

This normalised corpus, however, is not only interesting for automated processing but also for lexical research. The simple first person pronoun *ich* ('I'), e.g. knows many different spellings in the Swiss German dialect: *i, ich, ech, ii, y, iich, ig*, to mention only some of the variants found in the corpus. Additionally, the pronoun is often realised as a clitic: *hani* ('have I'), *sötti* ('should I') etc. The normalised corpus allows for a query of the standard spelling form ich and gives all the occurrences of all the different dialectal forms as a result.

The path from a non-standard SMS corpus to a normalised corpus and a PoS tagging is steep and involves many steps. Some of them were performed automatically, most of them manually by student helpers. In the following they are summarized for the corpus as a whole with examples from the individual languages. For more details per language, we refer you to the instructions written for the student helpers and written in the original language for

French

and

Italian

and written in German for

Romansh

. For German, similar rules were applied.

When applying the normalization, we used a self-developed software, which allowed for each student assistant to work on their own data while using a common database for the normalized data that had already been applied by any of the other student assistants. By using this approach, when student assistant A was about to annotate a type that had already been annotated in another dataset by student assistant B, student assistant A would receive a suggestion that showed student assistant B's approach when normalizing this type. By applying this system, we were able to ensure consistency within the normalization performed by different people. For more information about the mechanics of the steps performed, i.e. about the software used etc., we refer you to Ruef/Ueberwasser 2013.

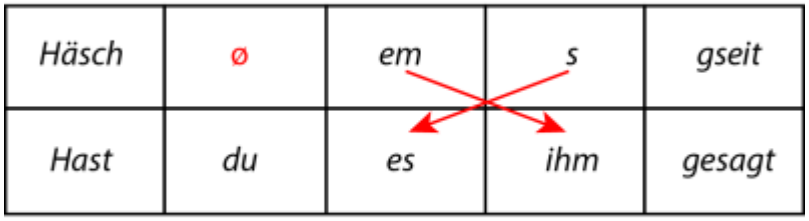
If you want to apply our methodology or quote from this documentation, please cite the documentation as noted in the bibliography.

General rules observed for all languages during the normalization

In an attempt to change as little as possible while still making the normalised layer available for research and processing, the following general rules were applied to all languages:

Syntax

No changes were made to the syntax, meaning that ellipsis were not completed and word order was not changed. The following example is perfectly grammatical Swiss German dialect ('have you told it to him?'), but in Standard German, the word order is different and the subject must be realised:
Swiss German dialect: *Häsch ems gseit?*



Standard German: *Hast du es ihm gesagt?*

This example thus is normalised as Hast ihm es gesagt? Non lexical elements

Elements that do not represent words (i.e. emoticons, punctuation) are taken into the normalized layer as they are, but they are annotated as <emo> </emo> <pun>!</pun> etc. in a special layer. This annotation was performed automatically where ever possible and then corrected by the student assistants. Elements that cannot be identified

In SMS texts, you find lots of words that are not understandable, because they represent family-lects, they are interjections or tokens in unknown languages or abbreviations that cannot be identified (e.g. tkdn). These elements were left as they appeared, i.e. they were taken over into the normalised layer in their original form. Foreign material

If tokens come from a foreign language that can be expected as known (i.e. Western European language) the spelling was adjusted to the original spelling where possible. If this token appeared in an inflected form, the inflection was added to the corrected spelling. kuul –> cool en kuuli idee –> eine coole Idee One to many or many to one

As has been said above, some tokens from the SMS layer had to be pulled together to present only one token in the normalised layer (e.g. [pomme] [de] [terre] –> [pomme de terre]), while others had to be taken apart ([hani] –> [han] [i]). These steps were noted down by our student assistants and executed by the computational linguists. Reconstructions and interpretations

We decided not to reconstruct or interpret language where it is not clearly recognizable. This goes especially for the case system in the Swiss German dialect and thus we use an according example here. In Standard German, at least in the masculine form, a nominative can be distinguished from an accusative based on the morphology (cf. den Mann vs der Mann in the following example). In the Swiss German dialect, on the other hand, this distinction is not visible (cf. dr Maa in both situations in the following example). Now, do the Swiss in fact use a nominative as an object or is this just an accusative that is homophone to the nominative? We don't know and - more importantly - we do not want to interpret. Accordingly, we decided to leave the noun and the article in the (standardized to der Mann) nominative in this situation. Case.png In a parallel approach, we do not reconstruct or interpret language in other situations or languages neither. Capitalisation

In German, nouns are spelled with a starting upper case letter. Independent of the capitalization in the SMS layer, nouns are in upper case in the normalized layer in an attempt to support a PoS tagger

in recognizing nouns. Spelling

Assuming that spelling is unorthodox in SMS all over, we decided to adjust spelling to what is found in a dictionary for a lemma on an according syntactical position. In German, e.g., there is an definite neutral article das and the conjunction dass (e.g. er sagte, dass er komme ('he said that he would come')). Irrespective of the spelling used in the SMS, we applied das for an article and dass for a conjunction. The same rule was applied for other homophonous words, too. Abbreviations

When abbreviations are used in SMS, three different situations can be distinguished: The abbreviation cannot be decoded, eg. tkdn. In this case, the abbreviation is taken over from the SMS layer into the normalized layer as is and it is marked as an abbreviation. The abbreviation can be decoded, e.g. cu for see you. In this case, the abbreviation is decomposed in the normalized layer, e.g cu becomes see you. This type of abbreviation is kept in the language, in which it is abbreviated, ie. see you is used in English even in otherwise German or French SMS. Abbreviations that stand for a brand or other type of name (e.g. IBM) were kept as they are. Digits

Digits were not modified, i.e. 3 remained 3 and three remained three. There is, however, one exception to this rule. Where digits were combined with letters, they were written out in the normalization, thus, 4tel became Viertel ('quarter'). Special rules for Swiss German dialect

Helvetisms

Helvetisms, i.e. lemmas that belong to Standard German in Switzerland according to the Variantenwörterbuch were taken over into the normalized layer in their standardized spelling, even though they might not be understandable to a reader from Northern Germany. No equivalent in standard German

Some words in Swiss German dialect do not have equivalents in Standard German, e.g luege ('to look') or gumpe ('to jump'). Where ever possible, we used lemmas with a similar sound to replace these expressions in the normalized level, provided the semantics are somehow similar. Following this idea, luege was transcribed as lugen (according to the Duden: "rural for to pry "). Where this type of approach was not possible, we normalized to the Standard lemma that is closest in its meaning, e.g. gumpe became springen. A special situation in this context is a verbal particle that can be realized as go, ga, goge and similar forms. This particle is syntactically compulsory in the dialect but has no equivalent in Standard German and is semantically empty. We decided to normalize this particle to go and to take it over into the normalized layer in this form. Prepositions

Quite regularly, the Swiss German dialect does not use the same prepositions as Standard German. In this case, we used the same preposition in the normalized layer as in the SMS layer (albeit adjusted in spelling where needed). E.g. i gane uf Bärn ('I go to Bern'), which should be ich gehe nach Bern in Standard German became ich gehe auf Bern. Diminutives

In Standard German a diminutive is normally realized as -chen, while the dialect only know a diminutive in -li. For some lemmas and in some (older) variants of German, a -lein diminutive exist(ed). Accordingly, we decided to apply this -lein form whenever a diminutive was used in the SMS. E.g. s'chindli ('the little child') became das Kindlein even though it sounds slightly archaic. Imperatives

In Standard German, the verb of an imperative can take a short or a long form: schlaf gut vs. schlafe gut. For the dialect, this is not the case, there is only a short form. Accordingly, the normalization always uses the short normalized form. Special rules for other languages

You find more information for languages other than German in the documaentations written in the

original languages for French and Italian and written in German for Romansh.

From:
https://sms.linguistik.uzh.ch/ -

Permanent link:
https://sms.linguistik.uzh.ch/03_processing/05_normalization?rev=1641298306

Last update: **2022/06/27 07:21**