Language tagging

Types of taggings

Each SMS was tagged for the languages contained within. There are three possible tags:

- Main language: For each SMS, a main language was defined as the dominant language, i.e. the language which provides most words to the SMS.
- Borrowing: Words from a language other than the main language. The words in the foreign language, however, have to be an established part of the main language's vocabulary.
- Nonce Borrowing: a word from another language than the main language of the utterance, which has not become an established part of this language (at the time of writing).

Restrictions

Because whole SMS were tagged rather than individual words, a few restrictions apply:

- You can search for SMS containing e.g. Italian words, but you cannot search for Italian words. Once you have the list of SMS that contain Italian words, you will have to find out by yourself, which words are Italian. The version of the corpus used in ANNIS, on the other hand, knows fewer languages but actually annotates words.
- You cannot search for the number of Italian words within the corpus, but rather for the number of SMS containing Italian words.

Identifying languages

To define, whether a word is an established part of the main language, the following codices were used for the specific languages:

- German: Duden Rechtschreibung as of 2010
- English: OED Online as of 2010
- French: Larousse français-monolingue as of 2010
- Italian: Garzanti Linguistica as of 2010
- Spanish: Diccionario de la lengua española as of 2010
- Swiss German: in order to distinguish between a dialectal expression and one that is established in the Standard specific to Switzerland), the following reference was used: Ammon, Ulrich et al. (2004): Variantenwörterbuch.

Normally, each SMS consist of exactly one main language. It can, however contain borrowings as well as nonce borrowings and both of them from different languages. In rare cases, an SMS can consist of two main languages. Look at the following example:

1. Olla fratello!!! Come stai? Wie geht's dir so? Immer noch so lange am arbeiten wie früher? Ich hab endlich mein eigenes Restaurant und mucho travajo aber macht mir extrem spass allora amore, buona giornata und luegsch uf di, gäll peace

This original SMS was tagged as follows:

- Main language: German Standard (because most words are in German Standard)
- Borrowings: French (because *Restaurant* is originally French but can be found in the Duden
- Nonce Borrowings:
    - Spanish (neither *travajo* nor *olla* can be found in the Duden). In spite of the unorthodox spelling, we consider both words to be Spanish, because no other language used in this SMS provides similar phonological variants.
    - Italian (the following cannot be found in the Duden: *fratello, come, stai, allora, amore, buona, giornata*)
    - English (*peace* cannot be found in the Duden)
    - Swiss German dialect (*luegsch, uf, di, gäll* cannot be found in the Duden)

As can be seen from this example, Standard German and Swiss German Dialect are tagged independently, because most likely a lot of research in this field will be performed on this corpus. The same goes for all Swiss national languages (i.e. German, French, Italian, Romansh), where we differentiated between varieties. However, varieties in other languages (e.g. British and American English or Spanish from Spain and Spanish from South-American countries) were not taken into consideration.

Special tagging problems

Swiss German dialect vs. Standard German

Since there is no fixed spelling system nor codex for the Swiss German dialect, many words become homograph in Swiss German and in Standard German. The following rules have thus been fixed to assign a word to one or the other variety:

- In case of homography, a word is considered to be the same as the main language. *ich* in a dialectal SMS is thus considered to be dialect, while ich in a non-dialectal SMS is considered to be Standard.
- If a seemingly dialectal word appears in a Standard SMS, the Variantenwörterbuch was consulted to differentiate between Standard Swiss German ('Helvetism') and dialect. If the word can be found in the Variantenwörterbuch, it is Standard, otherwise it is dialect.
- If a word, that sound like Standard German appears in a Dialect SMS, the person tagging the SMS (all native speakers of Swiss German) asked herself "could I use this in my own dialect?". In case of a positive answer, the word was considered to be Dialect, otherwise it was tagged as Standard.
- In some cases, it was not the individual word, but rather the word order that seemed to be Standard in an otherwise dialectal SMS. In this case we were very reluctant to register the Standard. It was only considered, if it is absolutely impossible to use the applied word-order in Standard Swiss German.
- Dialectal words in any other SMS (i.e. SMS in Standard German and in languages other than German) are considered to be nonce borrowings rather than borrowings, because a Dialect word never occurs in a codex.
- If a foreign word appears in a dialectal SMS, the Duden is used to find out, whether this word is a borrowing or a nonce borrowing.
- In some cases, a word could not be found in the Duden, but we did not consider it to be Swiss German dialect either. These tokens were marked as Other German, meaning they are dialectal

forms from Germany or Austria. The same goes for words that are to be found in the Duden but are marked as dialectal forms from Germany or Austria, e.g. the word *Bussi* ('Kiss'). It has to be noted that the decision to mark a token as Other German is partly based on the Duden but partly on the Swiss German native-speaker intuition of the people performing the language tagging.

Internationalisms

Some words occur in different languages (e.g. OK or *restaurant*). Many of them derive from Greek or Latin, these two languages where thus ignored as donors. They were actually ignored, even if the word was only coined in the last two hundred years or so. Photograph, e.g. was coined by the inventors of the camera and could thus be considered to be German or French depending on the sources. However, since we care about the word only and not about the coining, that type of word creation was ignored altogether, instead, the word is considered to be Greek and thus ignored. In all other cases, i.e. if the word is not a borrowing from Greek or Latin, the following rules were applied to define a word as a borrowing:

- Codices: If an internationalism is marked as e.g. being borrowed from English in the Duden, it was considered a borrowing.
- Pronunciation: If a word's origin is not marked in the codex but it still follows the pronunciation rules of a donor language, it was still considered a borrowing. E.g. the German verb *joggen* is clearly pronounced following the English rules, it is thus considered an English borrowing even though it is not marked as such in the Duden Rechtschreibung (n.b.: in most cases, the foreign origin was in fact marked in the Fremdwörter-Duden, however, that codex was not consulted by default).

Pseudo borrowings

There are words in a language, that clearly sound as if they were a borrowing, but are not known in the apparent donor language. An example would be handy, the German word for 'cell phone' or alles paletti (~ 'OK'), which imitates an non existent Italian word. These words are linguistically no borrowings. However, in the interest of possible research questions, we still considered them to be nonce borrowings.

Abbreviations

Abbreviations like *tgif* for 'thank god it's friday' where resolved whenever possible and thus considered for defining the main language (see below), but also to set borrowings or nonce borrowings. *tgif* would therefor in a German SMS be considered as an English nonce borrowing. On the other hand, abbreviations that could not be resolved such as *iLSi* were ignored. The following abbreviations were considered:

Token	**Interpretation**	**Language tagging based on:**
amt	Amo te	nonce borrowing: Portuguese
bb	bébé	only considered in French SMS
bjs	beijinhos	nonce borrowing: Portuguese
btw	by the way	nonce borrowing English

Token	Interpretation	Language tagging based on:
bmmw		bist mir mega wichtig
bx	bisous	nonce borrowing French
cu	see you	nonce borrowing English
fb	Facebook	ignored, since it's a proper name
Ga li gr		Ganz liebe Grüsse / Ganz liebi Grüessli
glg, GlG		Ganz liebe Grüsse / Ganz liebi Grüessli
Grz	Greetings (> greetz)	nonce borrowing English
hdl		Hab' dich lieb / Ha di lieb
lldvgHunvvvm buölugsnz vdbddadddkukdggf, KK		I lieb di vo ganzem Herze un no vill vill vill meh. [???] ...knuddel und küss di ganz ganz fescht, Kuss Kuss.
ka		keine Ahnung / Kei Ahnig
Kikoo	coucou	nonce borrowing French
ld		liebe dich / lieb di
ldsmf4iue	liebe dich so mega fest 4 immer und ewig	nonce borrowing English
lymtwcs	love you more than words can say	nonce borrowing English
lysm	Love you so much	nonce borrowing English
mdr	mot de rire	nonce borrowing French
piz	Peace	nonce borrowing English
t.b.c.	to be continued / confirmed	nonce borrowing English
tgif	thank god it's friday	nonce borrowing English
tk(...)		tausend Küsse (dein/e XY) / tuusig Küss
tqm	Te quiero mucho	nonce borrowing Spanish
tvtb	ti voglio tanto bene	nonce borrowing Italian
wdnv		will dich nicht verlieren / will di nöd verlüre
we	Wochenende / week-end	

Information used:

- http://linguistesblogueurs.blogspot.com/2005/11/les-abrviations-sms.html
- http://spanish.about.com/od/writtenspanish/a/sms.htm
- http://www.urbandictionary.com/

Phonetic writing

Phonetic writing, whether based on letters or on digits, occurs often in SMS. For the language tagging, we take their phonetic value and then tag them as normal words. E.g. *4* in an English SMS is treated as 'for'. If phonetic writing is ambivalent (e.g. *n8* can mean 'night' or 'nuit' or 'Nacht'), it is considered to be in the main language of the SMS. If, on the other hand *4* appears in a French SMS and cannot stand for 'quattre' but only for 'four' and thus phonetically for *for*, it is considered to be an English borrowing.

Unorthodox spelling

Spelling that deviates from prescribed spelling is not considered to make decisions about which language/variety a token comes from. This is especially important for the distinction between Standard German and dialect. Here, a /ß/ is not taken as a indication for an influence from Germany but rather as a typing variant on the mobile phone's keyboard.

Proper names

All kind of proper names were ignored when counting the tokens to decide on the main language. This includes brand names, people's names, names of places etc.

Homophony vs. homography

If a word's spelling implies a difference in pronunciation, the spelling is considered to define the language. E.g. *mam* in a German SMS is considered to be German, while *mom* is considered to be English because of the different pronunciation. If, however, the difference is only on the graphical level, it is neglected. E.g. *hey, hei, ey* are the same, in this case an English borrowing.

Compounds

To define the language of a compound consisting of elements from two different languages, the compound was taken apart and each part was treated individually. E.g. *Delegiertenmeeting* in an English SMS would have been tagged as a German nonce borrowing, because of *Delegierten-*. If the same compound had appeared in a French SMS, it would have been marked as containing each an English and a German nonce borrowing.

Main language

Defining the main language as the language with the most words sounds simple enough. However, it is not always that simple. The following problems occurred:

Abbreviations

SMS is a text type with lots of abbreviations, from *cu* ('see you') to *ihdmfg* ('i ha di mega fescht gärn'). When defining the main language, we counted words. Whenever possible, abbreviations where resolved and every word in the abbreviation (i.e. cu is two words) was counted individually. Thus, the SMS "Tgif : D btw, wrist roller drbi?" looked like German Dialect on the first look (roller can be German Dialect, drbi is definitely German Dialect). However, resolving those abbreviations, the SMS turned into "Thank god it [is] Friday : D by the way, wrist roller drbi" and is thus English.

Equal number of words

In case of equity (e.g. 2 words English, two words French), we tagged both languages as main language, because we want those SMS to be found by researchers of either language. In a case, where e.g. one word was clearly dialect, one word clearly Standard and all other words homograph, we decided the SMS to be Standard.

Compounds

If a compound deriving from two languages is needed to define the main language of an SMS, the compound is considered to be one word and the head is used to define the language of the whole word. E.g. in *Delegiertenmeeting* there is a German and an English component. However, since meeting is the head, the whole word is considered to be English.

No recognizable words

Some SMS come with only punctuation or emoticons. Their main language was tagged as 'unknown'.

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

Permanent link:
https://sms.linguistik.uzh.ch/03_processing/03_languages

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