Comparison of Automatic vs. Manual Language Identification in Multilingual Social Media Texts

Jennifer-Carmen FREY
Eurac Research, Italy

Egon W. STEMLE
Eurac Research, Italy

A.Seza DOĞRUÖZ
Independent Researcher

Multilingual speakers communicate in more than one language in daily life and on social media. In order to process or investigate multilingual communication, there is a need for language identification. This study compares the performance of human annotators with automatic ways of language identification on a multilingual (mainly German-Italian-English) social media corpus collected in South Tyrol, Italy. Our results indicate that humans and Natural Language Processing (NLP) systems follow their individual techniques to make a decision about multilingual text messages. This results in low agreement when different annotators or NLP systems execute the same task. In general, annotators agree with each other more than NLP systems. However, there is also variation in human agreement depending on the prior establishment of guidelines for the annotation task or not.

1. Introduction

1.1. *Languages in South Tyrol*

South Tyrol (ST, or Alto Adige) is a multilingual province in Northern Italy hosting German and Italian speakers and characterized by territorial and institutional multilingualism (Abel, Vettori & Forer, 2012). Italian, German and Ladin (a group of romance language dialects spoken in Northern Italy) are acknowledged as official languages and used in administrative communication and schooling. At the personal level, residents officially declare their language affiliation to ensure that public funding gets distributed according to the proportion of language groups[10]. Parents may choose the school and the language of instruction for their

[10] The latest census (2011) reports a proportional composition for the whole province of 69,6% of the population belonging to the German language group, 25,8% belonging to the Italian language group and 4,5% belonging to the Ladin language group. However, these numbers differ substantially in urban areas like Bozen/Bolzano and

children independently of their official language affiliation (e.g. parents, who declared that they belonged to the Italian language group may choose German as the language of education for their children or vice versa). Other languages spoken in the region are taught as foreign languages. In daily life, local residents (especially in urban areas) are in contact with speakers of other languages as neighbors, colleagues, classmates, or as friends and acquaintances. For example, language contact and multilingual communication occur as Italian loan words in the German dialect used in South Tyrol (DalNegro 2013), in code-mixing used in informal communication among mixed language groups (Ciccolone 2014) or in the multilingual "Corpus of South Tyrolean CMC Data" (DiDi corpus, cf. Section 2) (Frey, Glaznieks & Stemle 2016). Further information about multilingualism in the area can be found in Eichinger (2002), Dal Negro (2005), Forer (2010) and Abel, Vettori & Forer (2012).

1.2. *Multilingualism in online environments*

Multilingualism in online environments may refer to both the variety of languages in terms of web content and to the languages utilized by multilingual users across social media platforms on an individual level. For clarification purposes, we will discuss both briefly below.

Initially, most web content was created in English (Leppänen & Peuronen, 2012). With the increase of internet accessibility, the world wide web (WWW) has been filled with content in various other languages as well (Androutsopoulos, 2010; Danet & Herring, 2007, Tagg, 2015). The early assumption that English could take over all other languages through the developments of the WWW (Crystal, 2001; Paolillo, 2007) has changed with increasing proportions of non-English content (cf. Seargeant & Tagg, 2011; Hong, Convertino & Chi, 2011). This is mostly due to the user-generated content from social media platforms (Androutsopoulos, 2011b; Tagg, 2015). Social media contents are non-institutional, vernacular, interpersonal, (often) spontaneous and interaction-oriented instances of networked writing (Androutsopoulos, 2007). They encourage also smaller language communities to use their everyday language varieties and dialects (Androutsopoulos, 2011a). This certainly also contributes to the growing amount of multilingual contents in social media (Danet & Herring, 2003; Androutsopoulos, 2007; Leppänen & Peuronen, 2012).

Similar to other forms of daily communication, multilingual users communicate using multiple languages on social media platforms (see Coats, this volume). Freedom in writing, spelling and choice of language increases the amount of different languages used within one platform, even within one message, and fosters code-switching in blogs, status updates, comments, chat messages (Androutsopoulos, 2013b). Hong, Convertino & Chi (2011) studied multilingual and mixed language content on Twitter, Androutsopoulos (2013a), Belling & de Bres (2014), Schreiber (2015), Tagg (2014) on Facebook (FB), Leppänen & Häkkinen (2012) on YouTube and Leppänen (2007) on blog posts and web content. Frey, Glaznieks & Stemle (2015, 2016) and Glaznieks & Frey (2018) studied multilingualism in social media communication among the local residents of ST by compiling the DiDi corpus (cf. Section 2). The corpus contains both monolingual and multilingual messages donated voluntarily by the ST residents. Details about the data and users are explained in detail in section 2.

Meran/Merano, where Italian is spoken by the majority of the population (ca. 70%) and German has a smaller proportion (ca. 25%). The Ladin language is mainly spoken in the valleys of Badia and Gardena and is less present in other regions of South Tyrol (Astat 2016).

1.3. *Language identification in multilingual online environments*

Because of this fluid coexistence of different languages and varieties in online social media content, linguistic analyses depend on valid and fine-grained language identification of such texts. However, assigning language labels to multilingual social media texts is a tedious and complex task due to new and evolving writing conventions and the perception differences among annotators in terms of multilingualism. Moreover, with the increasing amount of data, manual analysis is often not feasible and the use of automatic methods for language identification becomes the only viable option for language annotations.

Apart from the linguistic community, there is a growing interest in analyzing multilingual and user-generated content automatically in the NLP community. Papalexakis & Doğruöz (2015) analyzed communication within multilingual social networks.

For most of the research in NLP, however, the main objective in language identification is to find ways to pre-process the data for later steps in the processing pipeline (e.g. named entity recognition, part-of-speech-tagging) as many steps depend highly on language-specific language models. There are a few automatic language identification systems that have been developed for multilingual texts. One of them is the polyglot[11] language identification system by Lui & Baldwin (2014). This system identifies all the occurring languages in a text (possibly from a predefined subset of all supported languages) and was specifically developed for multilingual social media texts. It is an experimental project that reuses the training data from the well-established tool langid.py developed for the monolingual setting (Lui & Baldwin, 2012). A probabilistic mixture model stipulates that each document is generated from an unknown mixture of languages from the training set. To select the set of languages that maximizes the posterior probability of the document, a Gibbs sampler first maps samples to languages.

The Compact Language Detector v3 (CLD3)[12] developed by Google is another common system for language identification to deal with multilingual texts. The CLD3 is an integral part of the Chrome browser (since 2016). In principle, it extends the influential and widely used approach for language identification from Cavnar & Trenkle (1994). While Cavnar and Trenkle (1994) consider sparse character n-grams for every language, CLD3 learns n-gram embeddings within a neural network architecture to model language identification.

Another recent and promising system is LanideNN by Kocmi and Bojar (2017) which uses single character embeddings within a neural network architecture. Using a dataset of artificially prepared multilingual documents (mixed from monolingual Wikipedia articles in 44 languages, average document length being ~ 5000 characters and containing 1-5 languages), Kocmi and Bojar (2017) report the following results for polyglot (Precision$_{\text{Macro}}$ .962, Recall$_{\text{Macro}}$ .963, F-measure$_{\text{Macro}}$. 961, Precision$_{\text{Micro}}$ .963, Recall$_{\text{Micro}}$ .964, F-measure$_{\text{Micro}}$ .963) and LanideNN (Precision$_{\text{Macro}}$ .962, Recall$_{\text{Macro}}$ .974, F-measure$_{\text{Macro}}$ .966, Precision$_{\text{Micro}}$ .954, Recall$_{\text{Micro}}$ .974, F-measure$_{\text{Micro}}$ .964).

Another attempt for automatic language identification for multilingual social media texts is the system developed by Nguyen & Doğruöz (2013). Classifying single tokens as one of the two languages used in a multilingual social media community, they report an accuracy value of .95 at the token-level.

In terms of evaluation studies for manual language identification tasks, King & Abney (2013) report a token-level percentage agreement value of 0.988 (with a chance level of 0.5) between two annotators for a manual annotation task of bilingual web documents. However,

[11] https://github.com/saffsd/polyglot

[12] https://github.com/google/cld3

the texts were not taken from social media and the annotators decided whether each token was in English or in another *a priori* known language.

Evaluating the performance of annotators on the data set used for the Tweet Language Identification Workshop at SEPLN 2014 (TweetLID) Zubiaga et al. (2016) report a percentage agreement of 92.6% between two annotators on an evaluation set of 3500 tweet messages. The tweets originated from Portugal and from the officially bilingual regions of Spain (e.g. Basque Country, Catalonia, and Galicia). They were mainly written in one or two of the six languages (Portuguese, the regional Languages, and Spanish and English). However, multilingual tweets are only about 6% of the whole corpus and 2% in the evaluation set. Inter-rater agreement drops to 60.9% when multilingual tweets are evaluated separately.

Although the results of all these studies are promising, they also reveal difficulties of processing multilingual social media texts automatically. In addition, there is need for more studies focusing on how humans perform on language identification tasks.

1.3.1. Method in Our Study

In our study, we have compared how humans (three in total) and automatic NLP systems (two in total) decide on a language identification task for multilingual (German, Italian, English) social media messages from the DiDi corpus.

We will first give a detailed overview on the data and the multilingual subset in section (2), present the tasks and procedures for human agreement and system evaluation in section (3), describe our evaluation methods in section (4), report and discuss our results in section (5), and conclude with a summary and suggestions for future work in section (6).

2. Data

In order to document and analyse multilingual language use in ST, in Northern Italy, the DiDi project (Digital Natives – Digital Immigrants. Writing on Social Networking Sites (SNS): A Corpus-based Observation of the current Language Use in South Tyrol, with Particular Consideration of the Writer's Age) collected social media texts by local residents, who voluntarily donated their messages (e.g. wall posts, wall comments and semi-private chat messages) published on the social networking site FB during 2013. The corpus aims at giving insights into non-institutional, undirected everyday language use of a supposedly multilingual individuals in an online environment and is searchable via the ANNIS corpus query infrastructure (https://commul.eurac.edu/annis/didi). It is also possible to download the corpus for scientific usage upon request. Below, we provide information about the corpus and describe how it was constructed. A detailed description of the corpus construction and its characteristics can be read in Frey, Glaznieks & Stemle (2015) and Frey, Glaznieks & Stemle (2016).

2.1. ***Data collection procedures***

Data donors were recruited via FB advertisements and announcements of the research project in regional FB groups and pages. To participate in the study, the donors received a link via FB to register and filled in an online questionnaire about their socio-demographic (gender, age, first language, dialect usage, employment status, internet usage habits and education of the participants) information. The study required an explicit user consent which also included information about privacy and terms of licence through a FB web application. After successful registration and completion of the questionnaire, an access link was generated via

FB and used subsequently to download the user's texts automatically. This procedure allowed for authentic, unaltered text messages, granted with explicit user consent and linked to the metadata of about the socio-economic backgrounds of the users. To speed up the recruitment process, we encouraged participants to advertise the web application and the research via FB within their social network. Nevertheless, the participation was fully voluntary and there was no further incentive for participants. One drawback of this effort is the bias on data collection through donors who were not chosen through sampling methods.

2.2. *Corpus construction and privacy measures*

The first step of the corpus construction was to access and to download the data (e.g. wall posts, comments and/or chat messages of the year 2013) that users agreed to donate via the FB Graph API[13]. Secondly, the user-related sociolinguistic metadata were linked to their texts. The texts were anonymized manually using generic tags to indicate the type of content replaced and sensitive personal information (e.g. user's name, other users in his/her network, hyperlinks, visual content) downloaded automatically from FB data streams was removed. The FB user id was replaced by an anonymous id to ensure the privacy of the participant but still make sure that data will be linked coherently. Comments and chat messages (from the corresponding threads) that were not written by the participant but automatically included in the FB Graph API download stream were also removed. The data were tokenized automatically using ark-twokenize-py (a Twitter specific tokenizer provided by: Myle Ott and based on Owoputi et al., 2013[14]) and corrected manually for the messages containing non-standard multi-language content. An initial language identification was done using langid.py (Lui & Baldwin, 2012) choosing the most probable language reported for each message. A manual investigation showed systematic issues with the automatic language identification for the social media messages in the corpus. Therefore, automatically assigned language tags were also corrected by humans afterwards, annotating the texts for the main language that can be observed. Frey et. al (2015) corrected all short messages (less than 30 characters), messages for which poor confidence levels (less than 0.8) were reported by langid.py, and messages that were tagged as any language other than German, Italian, English, French, Spanish or Portuguese. On the basis of the corrected language identification, a semi-automatic identification of multilingual texts was conducted to compare the out-of-vocabulary tokens with the dictionaries of other languages. During this step, texts where tagged as "multilingual" if out-of-vocabulary tokens in the text were found in the dictionaries provided by NLTK for other languages apart from the main language defined in the language identification beforehand. If out-of-vocabulary words did not appear in any of the dictionaries for German, English, Italian, French, Spanish or Portuguese, texts were tagged as "monolingual".

2.3. *Corpus size and overview*

The corpus comprises FB messages from 136 South Tyrolean Facebook users (63 male, 73 female). More specifically, there are 39,825 text samples (11,102 wall posts, 6507 comments to wall posts, 22,216 chat messages). 50 users allowed full profile access; 80 users allowed access to only semi-public (i.e. visible on their personal timeline and occurring in friend feeds depending on individual configurations (boyd, 2010)) posts and comments; and 6 people only to their private communication (i.e. chat messages). Data donors were between the ages of 14 and 76. 108 users declared German as their first language, 9 stated Italian, 2 Ladin and 1

[13] https://developers.facebook.com/docs/graph-api

[14] https://github.com/myleott/ark-twokenize-py

French. 16 users (11 German/Italian, 2 German/Ladin, 2 German/English and 1 German/Norwegian) named more than one language as their first languages simultaneously. The actual languages occurring in the corpus were highly influenced by the language biographies of individual users. Most messages were written in the first language of the user. As a consequence, there was a high number of German messages (58%) and Italian messages (21%), followed by English messages, which comprised 10% of the corpus. Only 1% of the corpus was written in other languages including French, Spanish, Portuguese, Japanese, Ladin or Latin. The rest (10%) of the messages did not contain any specific language content (emoticons, special characters, links, etc.) or they were entirely written with international phrases like *hi*, or *super*. Only about 7% of all the messages were tagged as multilingual in the semi-automatic identification. A detailed description of the languages found in the messages of the corpus can be found in Frey, Glaznieks & Stemle (2016).

2.4. *Subcorpus for study*

For the current study, we filtered the DiDi corpus (39,825 posts in total) as follows:

- only texts between 7 and 50 tokens (min, max) were included;
- duplicated messages would only appear once.

This filtering process eliminated 19,131 messages. Our sub-corpus consisted of 20,694 text messages (52% of the DiDi Corpus) from 81 users in total. We have randomly sampled text messages for the annotation task. In total, we extracted three text samples. The first batch of data consisted of 100 text messages and was only used for the annotators to become familiar with the type of data. The second and third batches of data consisted of 250 text messages. Each batch of sample data consisted of an equal amount of (allegedly) multilingual and monolingual text messages, making use of the annotations for 'multilinguality' provided with the DiDi corpus (cf. Frey et al. 2016) and described in section 2.2. The second and third batch consisted thus of 125, which were tagged as monolingual and 125, which were tagged as multilingual in the corpus. Hence, we artificially ensured a high ratio of multilingual texts in the test subset. This enabled us to highlight the difficulties of processing multilingual data without starting from a high baseline as observable for monolingual language identification.

3. Method

Since automatic systems have difficulties in processing multilingual social media texts, we assign the same task to human annotators for manual analyses and language identification. Below is a description of the procedure, information about the human annotators and the used automatic systems.

3.1. *Procedure for language identification using automatic systems*

For the automatic language identification, we selected two automatic language identification systems, the polyglot language identification for multilingual texts (Lui & Baldwin 2014) and the Compact Language Detector v3 (CLD3) developed by Google, and ran them on unprocessed plain text versions of the messages. Both systems provided a list of languages deemed to be present in the messages.We did not include LanideNN in our study due to time-restrictions. However, since the reported improvement compared to polyglot was rather low, we presume that polyglot gives a good estimation of state-of-the-art results.

3.2. *Human annotations*

Three human annotators manually identified the languages occurring in three batches (100 messages, 250 messages, 250 messages) of randomly selected text messages described in section 2.4. The annotators were linguists with experience in multilingualism and multilingual data. Two of the annotators are native speakers of German (from Germany and Austria) and have some understanding of Italian due to having lived in the ST area for several years. The third annotator has a passive knowledge of German and has not lived in the ST area. All annotators speak English fluently.

The first batch of data (100 text messages) and was used for the annotators to become familiar with the type of data. The second batch was annotated according to the task description “For each message in the sample corpus, list all the languages occurring in the message.” Besides this general goal, there were no guidelines and the annotators carried out their tasks individually without any communication or time restrictions.

After evaluating the results of the second batch, the annotators agreed on stricter guidelines for annotation defining concrete reference points (i.e. dictionaries for the most common languages) and a list of examples, which explained how to deal with problematic cases. The third batch was annotated manually following the established guidelines.

3.3. *Established annotation guidelines and annotation examples*

The guidelines for the second annotation round (third batch) maintained the general goal to list all the languages occurring in the message. Furthermore, the annotators agreed that other languages would only be added, when a word did not exist in the lexicon(s) of the language(s) that have already been assigned to the text message. In case of doubt, annotators referred to the well-known online dictionaries (e.g. Duden (https://www.duden.de/) for German (DE), Collins (https://www.collinsdictionary .com/) for English (EN), Olivetti (https://www.dizionario-italiano.it/) for Italian (IT) to check if the encountered words/phrases appeared in the dictionaries of those languages.

Named entities, which were problematic in the first annotation round (second batch), were treated as language independent and were ignored if they contained any foreign language content in the second round. Further problematic cases in the first round comprised loan words and loan word derivatives, intra-word-switches and interjections and greetings. Below are some examples from the data to illustrate the variety of multilingual text messages in our sample and how human annotators assigned languages to them according to the established annotation guidelines. Words in italics are in German, words in bold are in Italian and the rest are in English.

As a general rule, messages with any type of hybrid language usage (e.g. foreign language insertions, forms of code-switching or code-mixing, translations of contents) were annotated with all the languages occurring in the message:

Example (1)

User (1):	*Danke, danke fir olle Gratulationen* :) **Grazie mille a tutti per i congratulazioni**.	DE + IT
Trans:	Thank you for all congratulations :) Thank you all for the congratulations.	

Example (2)

User (2)*:*	sorry for that. *kann ich was für dich tun*?	EN + DE
Trans:	sorry for that. Can I do something for you?	

Both examples (1) and (2) are bilingual messages and were annotated as German and Italian (1), and English and German (2).

Loan words were only considered part of the language of the rest of the message if they appeared in the dictionary of the respective language (or of these languages, in case there are more visible and distinguishable languages). If loan words were not in the lexicon of the main language(s), the main language(s) of the message and the language of the loan word were listed.

Example (3)

User (3)*:*	*Ich denke nicht :/ sorry ich weiß noch nicht ob ich* <GeoNE> *fahre* :/	DE
Trans:	I don’t think so :/ sorry, I don’t know yet if I go <GeoNE>[15] :/	

Example (3) is annotated as only German. The word “sorry” is widely used among native German speakers and is listed in the dictionary (duden.de) for German.

Example (4)

User (4)*:*	like a *bössin* :P	EN + DE
Trans:	like a (female) boss :P	

Example (4) illustrates a creative usage of the German suffix "*-in*" that is usually used to indicate the gender of a person. In this case, it indicates that a female person did something like a boss. The message is annotated as English and German.

Interjections and greetings were treated like all other words, according to their existence or non-existence in the dictionary. If they did not appear in any of the dictionaries, they were treated as language independent.

Example (5)

User (5)*:*	**madai,** *warum muasch du mir des iaz vermiesen??* **Uffa** :(	DE + IT
Trans:	***oh,*** *why do you have to destroy this for* me?? **Sigh** :(	

In example (5), the user expresses emotions using the Italian interjections **madai** and **uffa** that are not listed in the German dictionary. Therefore, the message is considered as a mix between German and Italian.

Example (6):

[15] <GeoNE> was used to replace geographical named entities that have been removed for anonymisation purposes.

User (6)*:*	*ciao, jetzt muss ich los!...langsam.. da Schnee ist*	DE
Trans:	bye, no I have to go!...slowly..because there is snow	

“ciao” in example (6) is a valid German greeting according to the German dictionary (duden.de). Therefore, the message is considered as exclusively German.

Example (7)

User (7)*:*	*Für die Fans von Stanley Kubricks "Shining": der Dokumentarfilm "Room 237", ein Film über den Film, versucht die Rätsel dieses Schreckensfilms zu lösen. <PersNE>.*	DE
Trans:	For the fans of Stanley Kubrick’s “Shining”: the documentary “Room 237”, a movie about the movie, tries to solve the mysteries of this horror movie. <PersNE>[16].	

In example (7), “fans” is an English loan word that appears in the online German dictionary. Named entities (e.g. “shining”, “stanley kubrick”) are ignored in lines with the annotation procedure. As a result, the message is annotated as German only.

Example (8)

User (7)*:*	**No, a me l'unica cosa che non funziona, ma che aggiorneranno presto sono dei plugin di illustrator. Il resto va alla grande.**	IT
Trans:	No, for me the only thing that does not work but will be fixed soon, are the plugin of illustrator. The rest works perfectly.	

In example (8), “plugin” is an English loan word appearing in the Italian dictionary. “Illustrator” refers to a product by a brand. Therefore, its language label was ignored as a named entity. The message was annotated as Italian only.

4. Methodology for evaluation

We report Artstein & Poesio's (2008) Fleiss' Multi-π Kappa (where appropriate category-wise) and percentage agreement values for both annotation rounds, both with 95% confidence intervals approximated by bootstrapping (BCa, r=5000, n=250), and Cohen's d effect sizes for the difference between the two rounds. The data are considered reliable if human annotators agree on the assignment of established categories into units that are agreed upon. “[... However, if] the annotators are not consistent then either some of them are wrong or else the annotation scheme is inappropriate for the data. [... Conversely,] it is important to keep in mind that achieving good agreement cannot ensure validity" (Artstein and Poesio, 2008).

[16] <PersNE> was used to replace the named entities related to a person’s name that have been removed for anonymization purposes.

While percentage agreement simply reports the agreement between pairs of annotators divided by the total number of pairwise ratings, the kappa coefficient takes into account how much agreement is expected by chance and quantifies agreement over and above chance with values between 0 and 1. Usually, values larger than 0.4 are considered to represent moderate agreement, values larger than 0.6 substantial agreement, and values larger than 0.8 are considered very good for most situations (cf. Artstein and Poesio (2008) for more information).

The 95% confidence interval (95%-CI) reflects an amount of possible error in the sample and provides a range of values that are likely to include the true value to be within the lower and upper bound with the specified level of confidence. Cohen's d effect size expresses the mean difference between two measures in standard deviation units (the range is -3.0 to 3.0, and the interpretation of effect size consequently varies by context). To clarify, a Cohen's d of 1 means that 84 % of the values from the second measure will be above the mean of the first measure, 62 % of the values from the two measures will overlap, and there is a 76 % chance that a value picked at random from the second measure will have a higher score than a value picked at random from the first measure. Furthermore, with a Cohen's d of 3, 100 % of the values from the second measure will be above the mean of the first measure, 13 % of the values from the two measures will overlap, and there is a 98 % chance that a value picked at random from the second measure will have a higher score than a value picked at random from the first measure.[17]

Bootstrapping is a resampling method. Instead of validating a model by using complementary subsets as it is known from cross validation, we estimate the precision of a statistic of a data sample by randomly resampling with replacement from this one data sample and analysing the distribution of the statistic of the resamples. In our case, we estimate the precision of two inter-rater agreement statistics. To this end[18], we randomly resampled annotation decisions from all previously annotated texts until we had 250 for each annotator, calculated the inter-rater agreement between them, repeated this 5000 times, and then used this bootstrap distribution to estimate the confidence intervals.

All calculations were completed in R (R Development Core Team, 2017). Inter-rater agreement calculations were conducted with the irr package (Gamer et. al, 2012) and bootstrapping was done with the boot package (Canty and Ripley, 2017).

5. Results and discussion

This section reports on our evaluation results for the agreement across human annotators, across automatic annotation methods and across humans and automatic methods in comparison.

5.1. *Inter-rater agreement across human annotators*

Table 1 shows the Fleiss' Multi-π Kappa values for three annotators for the 250 items of round (1) and the 250 items of round (2). The Kappa agreement between the annotators increased for the second annotation round. Besides, we can see that the category-wise agreement for certain labels of language combinations increased. The calculated effect size and the small overlap between confidence intervals suggest that this is a substantial shift. Likewise, the percentage agreement also increased. This result suggests that the consultation between the

[17] http://rpsychologist.com/d3/cohend/

[18] Effectively, we calculated the bias-corrected and accelerated (BCa) bootstrap interval which also takes into account the estimate for the original data and adjusts for skewness in the bootstrap distribution.

annotators after the first round has resolved conflicts and led the annotators to agree more with each other during the second round.

For example, in the first round, the annotators agreed the least (.35) when assigning the label English-German. The discussion after the first round revealed that this issue was mainly due to English loan words (e.g. "sorry"), which were perceived as English (i.e. foreign) by some annotators and as German (i.e. not from a different language) by others. To resolve these disagreements, the annotation guidelines for the second round contained clear instructions on how to deal with loan words, intra-word switches, trans-scripting (cf. Androutsopoulos 2013b), and others. Example (3) and (5) in section 2 illustrate the guidelines for such cases.

For both annotation rounds, the annotators agreed most in assigning the label English (only). This might be due to the annotators' overlapping agreement about English words and the relatively smaller amount of German and Italian loan words in English. Notice that the percentage agreement for 'en (only)' in the second round is 100%. This means that every time at least one annotator classified a message as English all the other annotators agreed with this decision.

Table 1: Inter-rater agreement across three human annotators.

	First Round			**Second Round**			Effect Size
		95% CI-	95% CI+		95% CI-	95% CI+	
	Fleiss' kappa						
Overall	0.63	0.57	0.69	0.73	0.67	0.78	-3.3
De	0.63			0.74			
de_en	0.35			0.61			
de_en_it	0.58			0.78			
de_it	0.52			0.57			
En	0.93			1.00			
en_it	0.55			0.78			
It	0.80			0.85			
	%-Agreement						
Overall	63.60	56.80	69.16	74.80	68.00	79.20	-3.9
de (only)	57.41	48.77	64.20	70.19	62.11	76.40	-3.3
en (only)	81.82	29.08	90.91	100.00	-	-	-2.2
it (only)	60.66	45.90	70.49	68.09	51.06	78.72	-1.1

5.2. *Inter-rater agreement across machines*

Table 2 shows the Fleiss' Multi-π Kappa values for the first and the second annotation round for the two automatic methods polyglot and CLD3. The two systems cannot agree on the languages in a multilingual setting. They perform at about the level of human annotators prior to discussing guidelines in the monolingual cases. As expected, there is no increase in performance between the two rounds. The difference is only due to the regular performance differences on different data. The largely overlapping confidence intervals also point to this interpretation.

Table 2: Inter-rater agreement between two machines

	First Round			**Second Round**			Effect Size
		95% CI-	95% CI+		95% CI-	95% CI+	
	Fleiss' Kappa						
overall	0.47	0.40	0.55	0.43	0.36	0.50	1.3
De	0.61			0.61			
de_en	-0.02			-0.02			
de_en_it	0.00			0.00			
de_it	-0.07			-0.04			
En	0.65			0.43			
en_it	-0.01			0.00			
It	0.83			0.72			
	%-Agreement						
overall	63.60	57.60	69.20	62.80	56.80	68.40	0.3

As we used bootstrapping to calculate confidence intervals, we also validated the bootstrapping method. We sampled 125+125 new messages from the DiDi corpus, tagged them automatically with the two systems described in section 3.1 and calculated Kappa and percentage agreement values. We repeated this process 1000 times on the actual data. Afterwards, we calculated the mean Kappa and percentage agreement values, and their 95% confidence intervals.

Table 3: Validation of the bootstrap method on 1000 annotations runs of actual data.

		95% CI-	95% CI+
Fleiss' Kappa (overall)	0.46	0.46	0.46
%-Agreement (overall)	63.66	63.49	63.84

Table 3 illustrates that the bootstrapped values align with the actual data. However, the results for the second round (cf. Table 2) show an unusually low performance of the automated systems. This suggests that the data was more difficult to label. Therefore, improvements of the human annotators (cf. Table 1) stand out even more.

5.3. *Agreement between human and automatic annotation*

Table 4 illustrates the agreement between the human and automatic annotations. The overall agreement between the best performing machine and the humans is still considerably low.

Table 4: All humans and the best machine (polyglot) over two annotation rounds.

	First Round			**Second Round**			Effect Size
		95% CI-	95% CI+		95% CI-	95% CI+	
Fleiss' Kappa (overall)	0.59	0.54	0.64	0.64	0.58	0.69	-1.8
%-Agreement (overall)	50.80	44.40	56.80	59.20	52.40	64.80	-2.7

6. Conclusion and future research

Language identification for texts with multiple languages is a challenging task for both humans and automatic systems. Possible annotations/labels vary widely depending on how multilingualism is perceived by individual human annotators or modelled in automatic systems. In addition, variation in the decision-making process for humans and systems influence the outcomes of an annotation task. Although language identification of texts with multiple languages is often mentioned in relation to code-switching, traditional code-switching is “only one possible reason to explain why a document contains multiple languages, and is actually one of the less common causes” as reported in King & Abner (2013) for their corpus.

Our evaluation showed that agreement between human annotators and automatic systems on a dataset of 250 authentic Facebook messages (50% monolingual and 50% containing multiple languages) is fairly poor (0.59 Fleiss’ Kappa for the first annotation round and 0.64 for the second). Also, the comparison of two different automatic systems with each other only yielded low Kappa values (0.47 in the first round and 0.43 in the second round). The agreement of human annotators increased substantially for the second round (from 0.63 Fleiss’ Kappa to 0.73 Fleiss’ Kappa) after clear annotation guidelines had been established and agreed upon. This shows that consistency in language identification is not given a priori and the outcomes are highly influenced by what the annotators or the developers of a system regards as multilingual text. For consistent and comparable language identification, both linguists and the developers of automatic language identification systems could take into account that texts with multiple languages appear in many forms and manifestations in social media. While code-switching and more fluid forms of hybrid language use have already been well-studied in linguistics (May, 2013), the linguistic perspective in NLP is often reduced to one among many, which hampers the use of automatic systems for linguistic studies. Also, the NLP community's need for large data sets to train language models[19] leads to an opportunistic attitude towards getting as much data as possible (i.e. combining multiple independent data sets into one to train algorithms) (cf. Nguyen et.al, 2017). However, employing language identification tools for detecting multilinguality in documents within a linguistic context has to be taken with a grain of salt. Without a clear agreement among the annotators or developers of a system about what constituents a multilingual text for a specific task, reliable language identification will be difficult to attain both for humans and systems.

[19] In case of neural network based algorithms this need even increases tremendously.

BIBLIOGRAPHIC REFERENCES

ABEL, A., VETTORI, C., & FORER, D. (2012), Learning the Neighbour's Language: The Many Challenges in Achieving a Real Multilingual Society. The Case of Second Language Acquisition in the Minority–Majority Context of South Tyrol, in European Center for Minority Issues, European Academy of Bozen/Bolzano (Eds.), European Yearbook of Minority Issues (Vol. 9), Leiden, Netherlands, Brill Academic Publishers, 2271–2303.

ANDROUTSOPOULOS, J. (2007), Bilingualism in the mass media and on the internet, in Heller, M. (Ed.), Bilingualism: A social approach, New York, Palgrave Macmillan, 207–230.

ANDROUTSOPOULOS, J. (2010), Multimodal – intertextuell – heteroglossisch: Sprach-Gestalten in „Web 2.0"-Umgebungen, in Deppermann, A., Linke, A. (Eds.), Sprache intermedial. Stimme und Schrift, Bild und Ton, Berlin, de Gruyter, 419–445.

ANDROUTSOPOULOS, J. (2011a), Language change and digital media: a review of conceptions and evidence, in Kristiansen, T., Coupland, N. (Eds.), Standard Languages and Language Standards in a Changing Europe, Novus Press, 145–159.

ANDROUTSOPOULOS, J. (2011b), From Variation to Heteroglossia in the Study of Computer-Mediated Discourse, in Thurlow, C., Mroczek, K. (Eds.), Digital Discourse: Language in the New Media, Oxford, Oxford University Press, 277–298.

ANDROUTSOPOULOS, J. (2013a), Networked multilingualism: Some language practices on Facebook and their implications, International Journal of Bilingualism, 19 (2), 185–205.

ANDROUTSOPOULOS, J. (2013b), Code-switching in computer-mediated communication, in Herring, S. C., Stein, D., Virtanen, T. (Eds.), Pragmatics of computer-mediated communication, Berlin/New York, Mouton de Gruyter, 659–686.

ARTSTEIN, R., & POESIO, M. (2008), Inter-coder agreement for computational linguistics, Computational Linguistics, 34 (4), 555–596.

BELLING, L., & DE BRES, J. (2014), Digital superdiversity in Luxembourg: The role of Luxembourgish in a multilingual Facebook group, Discourse, Context and Media, 4–5, 74–86.

BOYD, D. (2010), Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications, in Papacharissi, Z. (Ed.), Networked Self: Identity, Community, and Culture on Social Network Sites, Routledge, 39–58.

CANTY, A., & RIPLEY, B. (2017), boot: Bootstrap R (S-Plus) function., R Package Version 1.3-20.

CAVNAR, W. B., & TRENKLE, J. M. (1994), N-Gram-Based Text Categorization, in In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 161–175.

CICCOLONE, S. (2014), Classificare il code mixing: una reinterpretazione dei parametri di constituency del modello di Muysken, Linguistica e Filologia, *34*, 95–134.

CRYSTAL, D. (2001), Language and the Internet, Cambridge: Cambridge University Press.

DAL NEGRO, S. (2005), Small languages and small language communities, International Journal of the Sociology of Language, 174, 113–124.

DAL NEGRO, S. (2013), Il prestito verbale nel contatto italiano-tedesco, Atti Del Sodalizio Glottologico Milanese, *7*, 192–200.

DANET, B., & HERRING, S. C. (2003), Introduction: The multilingual internet, Journal of Computer-Mediated Communication, 9 (1).

DANET, B., & HERRING, S. C. (2007), Multilingualism on the Internet, in Hellinger, M., Pauwels, A. (Eds.), Handbook of Language and Communication: Diversity and Change. Handbook of Applied Linguistics (Vol. 9), De Gruyter Mouton, 553–592.

EICHINGER, L. M. (2002), South Tyrol: German and Italian in a changing world, Journal of Multilingual and Multicultural Development, 23 (1–2), 137–149.

FORER, D. (2010), Direct and Extended Cross-Group Contact in South Tyrol: Effects on Attitudes and Identification of "German" and "Italian" Students, University of Trento.

FREY, J.-C., GLAZNIEKS, A., & STEMLE, E. W. (2015), The DiDi Corpus of South Tyrolean CMC Data, in Second Workshop of the Natural Language Processing for Computer Mediated Communication/Social Media. Proceedings of the workshop. University of Duisburg-Essen. 29 September 2015, Essen, Germany, 1–6.

FREY, J.-C., GLAZNIEKS, A., & STEMLE, E. W. (2016), The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts, in Corazza, A., Montemagni, S., Seneraro, G. (Eds.), Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), 5-6 December 2016, Napoli, Torino, Academia University Press, 157–161.

GAMER, M., LEMON, J., FELLOWS, I., & SINGH, P. (2012), irr: Various coefficients of interrater reliability and agreement, R Package Version 0.84.

GLAZNIEKS, A., & FREY, J.-C. (2018), Dialekt als Norm? Zum Sprachgebrauch Südtiroler Jugendlicher auf Facebook, in Ziegler, A. (Ed.), Jugendsprachen. Aktuelle Perspektiven Internationaler Forschung, Berlin, Germany, De Gruyter.

HONG, L., CONVERTINO, G., & CHI, E. H. (2011), Language Matters in Twitter: A Large Scale Study, in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 518–521.

KING, B., & ABNEY, S. (2013), Labeling the languages of words in mixed-language documents using weakly supervised methods, in Vanderwende, L., Daumé, H., Kirchhoff, K. (Eds.), Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, The Association for Computational Linguistics, 1110–1119.

KOCMI, T., & BOJAR, O. (2017), Lanidenn: Multilingual language identification on character window, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 927–936.

LEPPÄNEN, S. (2007), Youth language in media contexts: Insights into the functions of English in Finland, World Englishes, 26 (2), 149–169.

LEPPÄNEN, S., & HÄKKINEN, A. (2012), Buffalaxed superdiversity: representations of the other on YouTube, Diversities, 14 (2), 17–33.

LEPPÄNEN, S., & PEURONEN, S. (2012), Multilingualism on the Internet, in Martin-Jones, M., Blackledge, A., Creese, A. (Eds.), The Routledge Handbook of Multilingualism, London / New York, Routledge, 384–402.

LUI, M., & BALDWIN, T. (2012), langid. py: An off-the-shelf language identification tool, in Proceedings of the ACL 2012 System Demonstrations, Stroudsburg, PA, USA, Association for Computational Linguistics, 25–30.

LUI, M., & BALDWIN, T. (2014), Accurate language identification of twitter messages, in Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL, Gothenburg, Sweden, Association for Computational Linguistics, 17–25.

MAY, S. (2013), The multilingual turn: Implications for SLA, TESOL, and bilingual education, Routledge.

NGUYEN, D.-P., & DOĞRUÖZ, A. S. (2013), Word level language identification in online multilingual communication, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Association for Computational Linguistics, 857–862.

NGUYEN, D.-P, DOĞRUÖZ, A.S., ROSÉ, C.P., de JONG, F. (2016). Computational Sociolinguistics: A Survey. Computational Linguistics, 42, 537-593.

OWOPUTI, O., O'CONNOR, B., DYER, C., GIMPEL, K., SCHNEIDER, N., & SMITH, N. A. (2013), Improved part-of-speech tagging for online conversational text with word clusters, in Proceedings of HLT-NAACL 2013, Association for Computational Linguistics, 380–390.

PAOLILLO, J. C. (2007), How much multilingualism? Language diversity on the Internet, in Danet, B., Herring, S. C. (Eds.), The multilingual Internet: Language, culture, and communication online, New York, Oxford University Press, 408–430.

PAPALEXAKIS, E., & DOĞRUÖZ, A. S. (2015), Understanding Multilingual Social Networks in online immigrant communities. WWW'15. MWA Workshop, Florence, Italy.

R CORE TEAM (2017), R: A language and environment for statistical computing, Vienna, Austria, R Foundation for Statistical Computing.

SCHREIBER, B. R. (2015), "I am what I am": Multilingual identity and digital translanguaging, Language Learning and Technology, 19 (3), 69–87.

SEARGEANT, P., & TAGG, C. (2011), English on the internet and a "post-varieties" approach to language, World Englishes, 30 (4), 496–514.

TAGG, C. (2014), Translanguaging as an addressivity strategy for identity and relational work on Facebook, in Proceedings of the Conference Superdiversity: theory, method and practice in an era of change held by IRiS, University of Birmingham 23rd – 25th June 2014, 23–25.

TAGG, C. (2015), Exploring Digital Communication: Language in Action, London / New York, Routledge.

ZUBIAGA, A., SAN VICENTE, I., GAMALLO, P., PICHEL, J. R., ALEGRIA, I., ARANBERRI, N., FRESNO, V. (2016), TweetLID: a benchmark for tweet language identification, Language Resources and Evaluation, 50 (4), 729–766.