Challenges of error annotation in native/non-native speaker chat

Sviatlana Höhn, Alain Pfeiffer, Eric Ras
Luxembourg Institute of Science and Technology
Esch/Alzette, Luxembourg
sviatlana.hoehn@list.lu, alain.pfeiffer@list.lu,
eric.ras@list.lu

Abstract

This work addresses challenges related to error annotation of conceptually oral learner language such as instant messaging. The analysis is based on a corpus of German longitudinal native/non-native speaker instant messaging dyadic conversations. We show that deviations from language standard in instant messaging can be caused not only by lack of knowledge or high production pace, but also by speakers' competence in selection of interactional resources for the regulation of the social closeness, leading in this case to two contradicting identities of the same speaker: a competent language learner and a competent instant messaging user. We discuss the consequences from the perspective of language understanding and automated error correction in chat.

1 Introduction

Automatic processing of learner language is of interest for applications supporting writing activities, computer-assisted language learning and human-machine communication with non-native speakers. Because statistical natural language processing tools are mainly trained on texts produced by native speakers, their accuracy is much lower if applied for textual data produced by non-native speakers who have not yet fully mastered the language in which they interact. An additional level of complexity of such an analysis is introduced if medially written but conceptually oral (Koch, 1994) learner language needs to be processed automatically. Examples of conceptually oral but medially written interaction are chat and instant messaging (IM).

Chat and instant messaging have been studied as an additional (curricular and extracurricular) resource for language learning (Fredriksson, 2012; Marques-Schäfer, 2013; Tudini, 2010). To support language learning in chat, automatic error recognition and correction may be required. In this article we focus on native/non-native speaker instant messaging communication from the perspective of the error annotation. The error analysis needs to serve two needs: (1) natural language understanding in chat with an artificial conversation partner that helps to practice conversation in a second language, and (2)automatic error corrections in chat with such an agent or an Artificial Conversational Companion (ACC) (Danilava et al., 2013). The ACC should not play a role of a teacher or a tutor, but rather behave like a more knowledgeable and helpful peer. Because of the oral character of chat formulated by the concept of conceptually oral language, we need to discover errors that are potentially addressable in a chat-based Conversation-for-Learning (Kasper, 2004) as recorded and compiled in the dataset used for this analysis. A Conversation-for-Learning brings together participants because of their linguistic statuses of native and non-native speakers and combines elements of an informal conversation and a language classroom (e.g. sub-dialogues focusing on linguistic matters such as error corrections, see (Höhn, 2016) for a detailed analysis of such sub-dialogues). This work aims at answering the following research question: *What are the challenges in error annotation in conceptually oral learner language?* This question may have different answers depending on the purpose of error annotation. For instance learner language understanding and automated error correction performed by an ACC during the talk with a language learner may infer different sets of constraints and requirements for error recognition and error annotation. In addition, this question may have different answers for different speech exchange systems (e.g. informal chat, Conversation-for-Learning and form-focused language classroom).

We use a publicly available corpus of longitudinal dialogues between German native speakers and advanced learners of German as a second language (Höhn, 2015). The speakers in the corpus and in all examples in this article are encoded with N for native speakers and L for non-native speakers (learners). We keep the original turn and speaker numbers as well as orthography. English translations are added in italics. In addition, we had access to the unpublished part of the data collection which includes retrospective interviews with each participant in order to get insights into their linguistic choices. We apply a state-of-the-art error annotation scheme for German to answer the research question. The annotation scheme was created for the FALKO corpus (Reznicek et al., 2012) and is highly reused by other corpora discussed in the next section. In addition, we apply methods of Conversation Analysis (Markee, 2000) to analyse participants' selection of interactional resources in chat, such as specific forms of orthography.

2 Learner corpora and error annotation

Error-annotation of a corpus assumes a non-ambiguous description of the deviations from the norm, and therefore, the norm itself. A creation of such a description may be even problematic for errors in spelling, morphology and syntax (Dickinson and Ragheb, 2015). Moreover, different annotators' interpretations lead to huge variation in annotation of errors in semantics, pragmatics, textual argumentation (Reznicek et al., 2013) and usage of specific non-native language forms (Tetreault and Chodorow, 2008). Multiple annotation schemes and error taxonomies have been proposed for learner corpora, for instance (Díaz-Negrillo and Domínguez, 2006; Reznicek et al., 2012). Because error taxonomies are language-specific, we focus only on error annotation in German learner corpora in this article.

The situation with German error-annotated learner corpora is that there is a very small number of corpora, and only a small part of them are publicly available. The website[1] "Learner Corpora around the World" lists in May 2016 only 11 German learner corpora, 10 medially written and 1 spoken. In addition, there are a few publications about German error-annotated corpora not mentioned on the web page. Table 1 provides an overview on German learner corpora of which we were aware of at the time of writing this article. The table includes only information about the German part for multilingual corpora LeaP (Gut, 2009) and MERLIN (Boyd et al., 2014).

The major conceptual work on the annotation scheme and error taxonomy was accomplished by the FALKO team (Reznicek et al., 2013; Reznicek et al., 2012) and frequently re-used by the followers (German part of MERLIN (Boyd et al., 2014), EAGLE (Boyd, 2010), WHiG (Krummes and Ensslin, 2014)). WHiG is part of FALKO but contains texts from native speakers of British English who are intermediate learners of German.

The error-annotation for the mentioned corpora was approached in the following ways. The LeKo-corpus was created earlier than FALKO by the same principal investigator (Lüdeling et al., 2010). The corpus is accessible though FALKO platform. The researchers elaborated an error taxonomy on a small learner corpus of 30 texts that were written manually and then re-typed to make the resources digitally available and analysable. The difficulties with error annotation that were faced by the annotators of the LeKo corpus were taken into account in the annotation definition phase for the FALKO corpus. Specifically, some of the errors can be tagged differently depending on the *target hypothesis* - how the learners' intention is interpreted by the annotator. Dealing with such ambiguities became an issue for learner corpus annotation.

A multilevel annotation was introduced in the FALKO corpus in order to deal with different target hypotheses (Lüdeling et al., 2005). The minimal, first target hypothesis (orig.: *Zielhypothese*) ZH1 aims at sentence normalisation and is limited to only orthography and morpho-syntax, it is expected to make the sentence or utterance "understandable" for NLP tools. The second target hypothesis ZH2 should address all other types of errors, like semantics, lexical choice, pragmatics and style (Reznicek et al., 2012).

An extension of FALKO annotation schema has been suggested in the EAGLE corpus of beginning learner German where error numbering was introduced to deal with overlapping errors (Boyd, 2010). Multiple target hypotheses were handled by setting a preference for the target hypothesis which minimises the number of annotated errors.

ALeSKo is a corpus of annotated essays of advanced learners of German with Chinese as L1

[1] https://www.uclouvain.be/en-cecl-lcworld.html, retrieved on 31 May 2016

Title	L1	GFL level	Data type	Size	Error-annotated	Available
ALeSKo	Chinese	Different	Written texts	43 texts	Partial	Yes
CLEG13	English	B-C	Written texts	731 texts	NA	Yes
FALKO	Many	Intermed. - advanced	Written texts	Under development	Yes	Yes
WHiG	English	B2	Written texts	279 texts	Yes	Yes
MERLIN		A1-C2	Written examinations	1033 texts	Yes	Yes
LeKo	Many	Different	Written texts	30 texts	Yes	Yes
LeaP	Many	Different	Speech	183 records of 2-20 min	No	Yes
EAGLE		Beginners	Online work book, essays	50 WB & 81 essays	Yes	Yes
LINCS	English	Intermed.-advanced	Written texts, longitudinal	Under development	NA	No
ADS	English	Beginner-intermed.	Threaded discussion, chat, essays, longitudinal	Under development	NA	No
Telecorp	English	Different	Email, IM, essays	1,5 mio words	No	No
deL1L2IM	Russian	Advanced	IM	52000 tokens	Partial	Yes

Table 1: German learner corpora in May 2016

(Zinsmeister and Breckle, 2012). The annotation contains manual marks of topological fields (fields and error marking), referential expressions (definiteness, specificity, target hypothesis) and Vorfeld use. The subject of the ALeSKo study was coherence in learner texts based on the annotation of syntactic, referential and discourse information. German-L1 part of the FALKO corpus were used for L1-L2 comparison. A specific focus of the annotation in ALeSKo lies on referential expressions (Breckle and Zinsmeister, 2010), which are also in general an important area of NLP research.

A specific feature of CLEG13 corpus is that it has a longitudinal core of texts produced by students from their first year to their final exams (Maden-Weinberger, 2015). The corpus is accessible through FALKO platform.

In contrast to the written resources described above, the LeaP corpus includes phonologically annotated speech recordings of German and English learners of German (Gut, 2009). The corpus includes readings of nonsense word lists, readings of a short story, retellings of the story and free interviews.

The corpus KoKo is part of the project Korpus Südtirol, and focuses on German as a first language learned in South Tirol by school pupils (Abel et al., 2014). The corpus of German emails posted to USENET users described in (Becker et al., 2003) consists of ca. 120 000 sentences. An error typology of orthographic, morphological, morphosyntactic, syntactic and syntactic-semantic errors was taken as a basis for the error-annotation, however, only 16 error types from the typology were used for the corpus annotation.

The deL1L2IM corpus used for this work contains 72 dialogues of the duration between 20 and 90 minutes produced by pairs of German native speakers and advanced non-native learners during multiple weeks of IM interaction. Error annotation was performed only for selected types of errors that have been corrected by native speakers. A systematic error annotation of the dataset has been left for a future study (Höhn, 2015).

As (Meurers, 2009) notes, the annotation of learner corpora is mainly focused on annotation of learner errors, however, annotation of linguistic categories in learner corpora is also of interest. To create stable models of learner language for statistical NLP tools, information on occurrences of linguistic categories and their dependencies is required. This need is approached by linguistic annotation of learner corpora, similar as it has been done for native-speaker language. Examples of lin-

guistic annotation in learner corpora are (Amaral and Meurers, 2009) who focused on tokenisation in Portuguese interlanguage, and (Díaz-Negrillo et al., 2010) addressing the problem of POS-tagging in interlanguage. Related to the annotation of conceptually oral language, the challenge of POS-annotation in chat language has been addressed by (Bartz et al., 2014). The concept of grammaticality is applied to approach problems of syntactic annotation in learner language in (Dickinson and Ragheb, 2015).

Most of the error-annotated corpora consist of argumentative essays, and the developed error taxonomy is good for error-tagging in essays, but needs further elaboration to be fitted for conceptually oral language like instant messaging exchange. In contrast to a writing assistance program that has to (ideally) identify and correct *every* error, only a small amount of all errors are usually corrected in an ordinary conversation. Even in a language classroom, not every error is corrected in a fluency context. Therefore, there is a need to distinguish errors that could be potentially corrected in a Conversation-for-Learning from those, which should not be addressed to.

3 Language standard, chat conventions and L2 errors

In chat data, some deviations from the standard German do not count as an error. Sometimes it is even explicitly negotiated by the participants that, for instance, writing everything small will be declared as correct. Therefore, in addition to the objective identification of linguistic errors (difference between the produced language and the language standard), chat language needs to be analysed through the lens of conventions that are valid for the specific communication medium (chat in this case) and accepted by the interaction participants. This means that it cannot be completely defined in advance for chat, what will be an accepted deviation covered by conventions and what will "count" as an error that could be corrected, for instance:

1. Quick typing: everything that speeds up the typing pace does not count as errors: ignore capital letters in sentence and noun beginning, sentence punctuation.

2. Expressivity: word stretches (we found *Tor* with 62 *O*'s in it), uppercase, special symbols, punctuation symbols, quotes and parentheses, as well as various combinations of all of them.

3. Minor misspellings: typos are not important.

4. Oral style: not every utterance is a full sentence, word order is similar to oral.

There are explicit negotiations of typing rules. In all sequences that we found, we observed the following:

1. If participants engage in negotiations of spelling conventions, such negotiations are always initiated by the native speaker.

2. Production pace and conceptual orality of the interaction are the reasons for deviations addressed in chat, but not a lack of knowledge.

3. Deviations from language standard for the purpose of expressivity are not perceived as errors by chat participants.

Participant's linguistic identities (native or non-native speaker) also play a role in the selection of the applied spelling rules. For instance, participant N01 (native speaker of German) saw himself in chat with learners as a role model with respect to orthography. This is analysable in his use of, for instance, capital letters at the beginning of the nouns and utterances. N01 explains in the retrospective interview that he tried to write in according to German standard

> weil ich gegenüber nicht-deutschen-muttersprachlern versuche, die deutsche sprache so gut wie möglich in wort und schrift zu verwenden.
> *because I am trying to use written and oral German language as good as I can in communication with Non-German native speakers.*

However, N01 uses lowercase-only spelling during the interviews as opposed to the standard-compliant spelling that he chose to use in chat with non-native speakers. Thus, the orthography in chat which N01 uses with different partners is *recipient-designed*. Orthography compliance becomes an *interactional resource* in chat.

4 Orthography and social closeness

The presence of a high number of deviations from the language standard in text chat has been usually explained by a pressure to type quickly and demand for a high production pace in CALL studies

(Loewen and Reissner, 2009). However, language learners report that they had (or took) their time to use additional resources (such as dictionaries) for dealing with trouble in comprehension and production. Hence, the production overhead necessary for a standard-conform language in chat might be caused by participant's understanding of their social roles of language novices and language experts, and used for the regulation of social closeness.

Example 4.1. Mutual dependencies between orthography and social closeness.

1	L03	Hallo! Entschuldigung, Ich weiß nicht, wie heißen Sie. Ich bitte um Verzeihung, ich habe total über heutige Unterhaltung vergessen. Ich schäme mich, wirklich, aber ich war beschäftigt, und musste dringend einige Probleme lösen, deshalb habe ich total über den Chat vergessen- ich bitte noch ein Mal um Entschuldigung, und verspreche, dass es nie wiederholen wird. Ich hoffe, dass unser Chat wird uns Spaß machen. mit freundlichem Gruß, L03_FirstName L03_LastName! *Hello! I am sorry, I don't know your [III p. pl.] name. Please forgive me, I totally forgot about [* error: wrong preposition] today's conversation. I feel ashamed, really, but I was busy, and had to solve several problems urgently, this is why I totally forgot about [* error: wrong preposition] the chat - please forgive me again, I promise that it will never happen again. I hope that our chat will be pleasant. best regards, L03_FirstName L03_LastName!*
2	N02	Hallo L03_FirstName, das ist überhaupt kein Problem! Ich hoffe, alle Probleme sind gelöst und wir können ein bisschen chatten. *Hello L03_FirstName, it is absolutely no problem! I hope, all the problems are solved and we can chat a little bit.*
3	L03	Ja, natürlich! wie heißt du? *Yes, of course! what is your [II p. sg.] name?*
4	N02	oh Entschuldigung, ich heiße N02_FirstName, bin 27 Jahre alt und wohne in München. *oh, I'm sorry, my name is N02_FirstName, I am 27 and live in Munich.*
5	L03	sehr angenehm! und ich bin 21 und wohne in Vitebsk, Belarus! *very pleasant! and I am 21 and live in Vitebsk, Belarus!* nice to meet you! and I am 21 and live in Vitebsk, Belarus
6	N02	oh, ich bin schon alt ;) *oh, I am already old [smile]*
7	N02	warst du schon mal in Deutschland? Ich war noch nie in Belarus *have you already been to Germany! I have never been to Belarus*
8	L03	ja, aber ich bin schon verheiretet))) *yes, but I am already married [smile]*
9	N02	oh echt?? wow! seit wann denn, wenn ich fragen darf? *oh really?? wow! may I ask you, how long?*

Example 4.1 presents the very beginning of the talk between L03 and N02. Because the participants have never met before, L03 does not know, who is on the other side of the connection. She comes too late to her first appointment and formulated her first message (turn 1) to her chat partner in a very polite way using a polite German form of address *Sie* (III p, pl., no English equivalent). In addition, she produces an email-like turn - conceptually closer to written than oral language - according to German spelling standard and closes if with a "best regards + name" untypical for instant messaging.

L03 produces multiple morpho-syntactic and semantic errors, however, her phrases start with a capital letter (except of the closing expression), and she is doing her best in positioning herself as a competent German speaker. N02 answers with a "no problem", and her message satisfies the German language standard, too. L03 switches from *Sie* to *du* (you, II p. sg.) in turn 3. In addition, she changes the spelling in the second phrase starting with a small letter instead of a capital. N02 responds with changed applied standard in turn 4 writing only nouns with an initial capital letter.

The participants continue with the rule "write only nouns with a capital letter". Shorter time intervals between turns 5-9 in Example 4.1 show how higher engagement leads to higher talk pace and therefore higher production pace. Deviations from language standard are the price for the typing pace, but in addition, they express a higher grade of engagement and social closeness.

There are mutual dependencies between participants' choices in terms of language standard. A closer look at the native speaker N02 and her partners L03, L04 and L05 helps to understand how participants deal with spelling and punctuation conventions, and how they influence each other. We discuss here only the results, the original data can be obtained from ELDA (Höhn, 2015). N02 behaves differently with her different partners:

L03 Both participants start with the standard-compliant spelling and shift then to a version where they move between standard-compliant spelling and "write-only-nouns-with-a-capital". L03 starts with *Sie* but switch to *du* in turn 3.

L04 starts with a "relaxed" version of spelling: only nouns are written with a capital, a very oral style. N02 starts with a norm-compliant version but adapts to non-native speaker's spelling version after ten turns. Later on, both participants even use lowercase for all

words. L04 starts with *du*. Overall chat of this pair can be characterised as very oral: short phrases, quick, many short turns.

L05 starts with a norm-compliant orthography and *Sie*. L05 makes lexical errors in her first turn. N03 replies with *Sie* but she decides to write the first word in each sentence small. Later on, L03 changes between a norm-compliant spelling and the relaxed "first-letter-small" version. L05 adopts this way of spelling from time to time. In the second chat, L03 start with *du* (first turn in this meeting) using proper spelling, but switches later to the relaxed "first-letter-small" version. It remains an open question if N03 noticed that L05 is not that much an independent language user (compared to the others) and shows her, how to do "chat-in-German".

The other native speakers in the dataset prefer to keep the same orthography style with all their partners: N01 presents himself as a role model, N03 prefers to optimise the spelling to increase the typing pace and types everything with lowercase, and N04 normally types all nouns with an initial capital, but starts all new sentences with a small letter.

5 Learner error annotation

In order to test the error taxonomy, we selected an initial set of data consisting of 481 questions produced by language learners. The error-annotation of the questions was performed according to the annotation guidelines for FALKO Corpus of German learner language (Reznicek et al., 2012; Reznicek et al., 2013). ZH1 was constructed according to the rules of standard German grammar and orthography with Duden dictionary as a reference. Semantics, lexical constructions and pragmatics are the subject of the extended, second target hypothesis ZH2. Example 5.1 shows the two target hypotheses for a sample question.

Example 5.1. Creating target hypotheses for error correction in questions.

402 L08	und um wieviel Uhr gehst gewöhnlich zum Bett?
	and at what time do you normally go to the bed?
ZH1	Und um wie viel Uhr gehst du gewöhnlich zum Bett?
	And at what time do you normally go to the bed?
ZH2	Und um wie viel Uhr gehst du normal ins Bett?
	And at what time do you normally go to bed?

The questions have been annotated by two human annotators: one German native speaker and one non-native speaker with a near-native level of German proficiency. Both annotators had sound background knowledge in corpus annotation and were experienced users of instant messaging. The following issues were faced when annotating errors in chat according to FALKO guidelines. First, special symbolic and orthographic means of expressivity used in chat must be classified as errors according to Duden and FALKO error annotation guidelines. Second, FALKO annotation guidelines do not provide any specific instruction for the cases where the errors in the verb make more than one target for the verb possible.

Example 5.2 illustrates one of the cases. This error has been corrected by the interaction partner of L09 in the dialogue and both possible targets for the erroneous question were addressed in the correction. Therefore, having in mind the application where corrections should be automatically generated in a conversation, we add both target hypotheses to the annotation.

The differences between the original learner's utterance and the two target hypotheses help to classify the errors and to generate corrections. In addition, it allows to analyse empirically what normalisation steps are really required for automated language understanding.

Example 5.2. Ambiguous target hypotheses.

135 L09	gefiel dir das studium leicht?
	Unclear target: Was the study easy for you? or Did you like your study?
ZH1a	Gefiel dir das Studium?
	Did you like your study?
ZH1b	Fiel dir das Studium leicht?
	Was the study easy for you?
136 N04	es gefiel mir, aber es fiel mir nicht immer leicht ;-)
	I liked it but it was not always easy for me
137 N04	("gefallen" = "etwas schön finden",
	("to like" - "to find something pretty",
138 N04	etwas fällt jemandem leicht = man hat keine Mühe damit)
	something is easy for someone = one has no effort with it)

However, chat conventions allow writing everything with small letters only and do not consider typos as errors that need a correction. This is why information about potential correctability of the errors in chat need to be encoded in the error annotation. Additional rules for exceptions need to be specified when deviations in orthography and punctuation are used as a means of expressivity. Therefore, we introduced the "real" error flag with the purpose to identify all errors that are *potentially addressable* in chat. The conventions that we take

into account for the "real" error flag are restricted to orthography and allow to:

1. start an utterance, a new sentence and nouns with a small letter,
2. write lowercase or uppercase or camel-case,
3. use punctuation and special symbols for the purpose of expressivity (emoticons),
4. omit punctuation and to use emoticons to separate turn-constructional units,
5. produce word stretches.

These rules are consciously applied by chat participants while typing. In addition, there are misspellings which are the result of a high typing pace and not lack of knowledge. They also do not qualify as errors in chat and are not considered as "real" errors. There are two exceptions that we annotate as real errors:

1. If a speaker repeats the same misspelling several times and the misspelled word sounds exactly as the correctly spelled word.
2. If it is a special, difficult case where even native speakers often make mistakes, for instance *ziemlich*.

With Duden as a reference for the language standard, 428 questions would contain an error. However, only 136 questions contained "real" errors. Only 21 of all potentially addressable errors in questions have been corrected in conversation by the native speakers. This low number of corrections is mainly explained by the type of the speech exchange system recorded in the corpus. An artificial conversation partner will need to decide in real-time, which of the potentially addressable errors may trigger a correction. This problem has been captured in a correction decision model and discussed in (Höhn, 2016).

As already reported in earlier academic publications, finding a target hypothesis may be a hard problem even for human annotators (Reznicek et al., 2013). We faced the same issue in our work. More specifically, ZH1 and ZH2 may correct different, mutually excluding errors.

Example 5.3 illustrates an error in plural in a non-native-like expression. ZH1 corrects only the error in the plural, confirming the use of the non-native-like expression. ZH2 corrects the non-native-like expression, but does not address the error in the number. Both errors can be hardly address by one target hypothesis.

Dealing with multiple target hypotheses will also be an issue for a computer program that should produce a correction. If only one target hypothesis should be presented in the correction, then criteria for selection need to be specified. As Example 5.2 shows, this is not always possible. If multiple target hypotheses can be presented to the user in a writing assistance program, then the user will have to choose one of them for the correction. This may be less helpful for the user if he or she does not have the necessary level of linguistic competence in the foreign language to make this choice. In both cases the program will need to guess what the user could have meant.

Example 5.3. Different target hypotheses correct different errors. Trouble sources are underlined. Target hypotheses are added in the bottom.

79	L03	)))) ja, wahrscheinlich! Sind die Grenze des Schuljahres von Urlaubssaison in Beiern abhängig? *yes, probably! Do the border of the school year [* errors: verb-subject number congruence, uncommon expression] depend on the holiday season in Beiern [* error: spelling]?* yes, probably! Do the borders of the school year depend on the holiday season in Bavaria?
80	L03	* Bayern * Bayern [self-correction] **Bavaria*
81	N02	ja genau! ist das bei euch auch so? *yes, exactly! is it like this in your place, too?*
	ZH1	Sind die Grenzen des Schuljahres von der Urlaubssaison in Bayern abhängig? *Do the borders of the school year depend on the holiday season in Bavaria?*
	ZH2	Sind die Ferienzeiten von der Urlaubssaison in Bayern abhängig? *Do the school holidays depend on the holiday season in Bavaria?*

6 Findings and discussion

Conceptually oral learner language such as instant messaging and chat introduces additional levels of complexity in error annotation as compared with conceptually written learner language (e.g. essays). In this work we make an attempt to discover these challenges on an example of a German native/non-native speaker instant messaging corpus deL1L2IM (Höhn, 2015).

Section 4 shows that participants of an instant messaging chat use deviations from language standard as a interactional resource to regulate social

closeness and to present themselves as members of specific social categories. Such categories are for instance a native speaker who positions himself as a role model, as well as a competent non-native speaker who is a competent instant messaging user. In this way, this work supports observations described in (Lee et al., 2012) that the concept of language correctness (or grammaticality in (Lee et al., 2012) is not an absolute category, but may vary depending on the level of formality or social proximity.

Using deviations form linguistic standard belongs to the interactional competence in computer-mediated communication and therefore, such deviation should not be always classified as errors. Such deviations in German native/non-native speaker chat mainly include orthography of German nouns and initial letters of an utterance, but also oral verb forms (*hab* instead of *habe* and more oral forms of question (e.g. declarative utterances which are functional questions, see also (Stivers and Enfield, 2010) for question classification).

In addition, the chat conventions covering such deviations (what is allowed) may vary for different pairs of speakers and change for the same pair of speakers with the time. This is the consequence of the variance in the level of social closeness for different pairs of speakers, and changes in the grade of social closeness that may occur with the time for one pair of speakers, since their relationship may change. Nevertheless, the set of potentially correctable errors seems to be quite stable for the specific speech exchange system (here Conversation-for-Learning). It was acceptable even for those native speakers who preferred typing according to German orthography standard, if learners typed with deviations in orthography (e.g. omission of initial capital letters). In addition, the types of deviation produced by the native speakers in the dataset differ from those produced by the learners. For instance, usage of oral forms of verbs (e.g. *hab* instead of *habe*, Engl.: *have* I p. sg) was only observed in utterances by native speakers. This may obscure learner's familiarity with oral German which is not explicitly covered in language classes or by language tests. Thus, some types of deviations may signal higher levels of familiarity with specific aspects of the foreign language use.

With this observations, the error annotation in conceptually oral learner language needs to cover at least one additional layer, namely the layer of potentially correctable errors in order to serve the need of automated error correction in conversation (although the occurrence of a potentially correctable error does not immediately trigger a correction, see (Schegloff, 1988)). We approached the problem of identification of potentially correctable errors by the "real" error flag although we also considered other possibilities which we discuss below.

One possible approach to identify such errors could be a comparison by the chat conventions applied by the native speakers in chat. However, this approach has at least two shortcomings. First, our data show that some native speakers may make their social roles of language experts more important than their roles of proficient IM users, and purposely avoid any deviations from language standard (see Sec. 3, example with N01). This pattern is not necessarily taken up by the learners. In this case, potentially correctable errors would include all those minor deviations that normally do not count as errors in a chat-based Conversation-for-Learning. Second, this approach would automatically put the native speaker in the position of a language expert, and the non-native speaker in the complimentary category of a language novice. However, being a native speaker of a language does not necessarily correlate with high language proficiency. This is why the notion of expertise or *differential language expertise* is suggested by the CA community as more appropriate to describe the socio-linguistic data in native/non-native speaker communication (Hosoda, 2006). For these reasons we suggest to analyse errors and deviations in learners' utterances independently from utterances of their native speaker partners.

Another way to identify correctable errors in IM chat would be looking at those errors that have been corrected by native speakers. The main limitation of this approach is that only a small number of errors received a correction in the dataset, and the number of corrected errors highly varied among different native speakers: some learners produced a high number of errors, but they were not corrected by their partners. An identification of a potentially correctable error in chat does not automatically mean the necessity of a correction, which is also confirmed by the numbers in our dataset (only 21 corrections of 136 "real" errors). Therefore, we relied on the intuitive concept of the "real" error in our analysis.

Deviations from language standard in chat may

occur because they are produced by instant messaging speakers consciously with the purpose of regulation of the social closeness. They can be also produced unconsciously due to high typing pace. The high typing pace, in turn, may be caused by the time pressure, but also by a high participants' engagement in talk. Lack of knowledge is rarely the reason for such deviations. However, it might be important for the language understanding components to find a normalised, grammatical equivalent to learner's utterance. The analysis of the set of learners' questions in Sec. 5 shows that the first target hypothesis ZH1 already serves this need for the analysed dataset. However, the effectiveness of ZH1 for this purpose may be different for less advanced language learners.

Section 5 also shows that learner errors make it sometimes necessary to consider several target hypotheses on each level (Example 5.2). In conceptually oral learner language, this needs to be done not only to guarantee the correctness, but also to maintain intersubjectivity and mutual understanding in the talk. Therefore, additional sub-levels in error annotation may be needed, as suggested in Example 5.2. These additional levels of error annotation can be used by the agent in real-time to capture multiple possible meanings of learner's utterance. Possible responses to such utterances include error corrections with disambiguation like in Example 5.2, or repair initiations (frequently called clarification requests in academic publications in NLP community, see for instance (Schlangen, 2004)).

7 Conclusions and future work

While FALCO annotation guidelines (Reznicek et al., 2012) already provide a comprehensive basis for error annotation in conceptually written learner language, annotating conceptually oral learner language brings the annotation task to a higher level of complexity. Specifically, there is a need to distinguish between deviations from language standard which can be addressed as an error in chat, and all other types of deviations which do not count as error in chat due to chat conventions (produced consciously or unconsciously).

The learners' level of proficiency in the foreign language influences the frequency of errors. However, a high level of familiarity with computer-mediated communication may lead to an increased number of deviations. This makes error annotation in conceptually oral learner language more difficult, namely the decision whether a deviation is caused by a lack of knowledge (and is potentially correctable) or by the competence in language use (and should not be addressed to).

Because chat conventions may change over time and may be different for different pairs of participants, a further question for research may be an automated recognition of the chat conventions and their incremental adaptation.

As argued in this article, orthography (or deviations from it) is an interactional resource in chat used by participants to regulate the social closeness. An open question remains, how these observations may be captured in a computational model for an artificial conversation partner or an ACC aiming at long-term interaction with the user (multiple weeks). Because all of the native speakers in the dataset show different behaviour with this regard, orthography as an interactional resource may be also a means for expression of specific characteristics of agent's individual interaction profile (Spranz-Fogasy, 2002; Höhn, 2016).

Because the identification of a potentially correctable error does not necessarily trigger a correction, one way for an ACC to handle uncertainties in error recognition is to decide against an error correction, first of all. Uncertainties in language understanding (caused by learner errors or other issues) can be either handled in the dialogue using repair practices or making use of *contingency* which is present in talk at virtually every point (Schegloff, 1996). Contingency allows to have more than one options for responses after each utterance, which makes dialogue modelling difficult but allows to introduce "back doors" in dialogue (types of turns that are valid next turns after the turn where an uncertainty with language understanding occurred).

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. KoKo: an L1 Learner Corpus for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2414–2421.

Luiz A. Amaral and W. Detmar Meurers. 2009. Little things with big effects: On the identification and interpretation of tokens for error diagnosis in icall. *CALICO Journal*, 26(3):580–591.

Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommu-

nikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Zeitschrift für germanistische Linguistik*, 28(1):157–198.

Markus Becker, Andrew Bredenkamp, Berthold Crysmann, and Judith Klein. 2003. Annotation of error types for German Newsgroup Corpus. In *Treebanks*, pages 89–100. Springer.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC*, pages 1281–1288.

Adriane Boyd. 2010. EAGLE: an Error-Annotated Corpus of Beginning Learner German. In *Proc. of LREC*. ELRA.

Margit Breckle and Heike Zinsmeister. 2010. Zur lernersprachlichen generierung referierender ausdrücke in argumentativen texten. *Textmuster: schulisch-universitär-kulturkontrastiv.*, pages 79–101.

Sviatlana Danilava, Stephan Busemann, Christoph Schommer, and Gudrun Ziegler. 2013. Towards Computational Models for a Long-term Interaction with an Artificial Conversational Companion. In *Proc. of ICAART'13*.

Ana Díaz-Negrillo and Jesús Fernández Domínguez. 2006. Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, 19:83–102.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. In *Language Forum*, volume 36, pages 139–154.

Markus Dickinson and Marwa Ragheb. 2015. On grammaticality in the syntactic annotation of learner language. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 158.

Christine Fredriksson. 2012. About collaboration, interaction, and the negotiation of meaning in synchronous written chats in l2-german. In Linda Bradley and Sylvie Thouësny, editors, *CALL: Using, Learning, Knowing, EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings*, pages 88–92. Research-publishing.net.

Ulrike Gut. 2009. *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*, volume 9 of *English Corpus Linguistics*. Peter Lang.

Sviatlana Höhn. 2015. deL1L2IM: Corpus of long-term instant messaging NS-NNS conversations. ELRA http://islrn.org/resources/339-799-085-669-8/.

Sviatlana Höhn. 2016. *Data-driven repair models for text chat with language learners*. Ph.D. thesis, University of Luxembourg.

Yuri Hosoda. 2006. Repair and relevance of differential language expertise in second language conversations. *Applied Linguistics*, 27(1):25–50.

Gabrielle Kasper. 2004. Participant Orientations in German Conversation-for-Learning. *The Modern Language Journal*, 88:551–567.

Peter Koch. 1994. Schriftlichkeit und sprache. In *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung*, pages 587–604. Walter de Gruyter.

Cédric Krummes and Astrid Ensslin. 2014. What's Hard in German? WHiG: a British learner corpus of German. *Corpora*, 9(2):191–205.

Sun-Hee Lee, Markus Dickinson, and Ross Israel. 2012. Developing learner corpus annotation for korean particle errors. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 129–133. Association for Computational Linguistics.

Shawn Loewen and Sophie Reissner. 2009. A comparison of incidental focus on form in the second language classroom and chatroom. *Computer Assisted Language Learning*, 22(2):101–114.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, pages 15–17.

Anke Lüdeling, Elena Briskina, Julia Hantschel, Jenny Krüger, Stéphanie Sigrist, and Ulrike Spieler, 2010. *LeKo Lernerkorpus Handbuch*. Humboldt-Universität zu Berlin, Institut für Deutsche Sprache und Linguistik, Philosophische Fakultät II.

Ursula Maden-Weinberger. 2015. "Hätte, wäre, wenn...": a pseudo-longitudinal study of subjunctives in the corpus of learner german (CLEG). *International Journal of Learner Corpus Research*, 1(1):25–57.

Numa Markee. 2000. *Conversation Analysis*. Mahwah, N.J.: Lawrence Erlbaum.

Gabriela Marques-Schäfer. 2013. *Deutsch lernen online. Eine Analyse interkultureller Aktionen im Chat*. Gunter Narr Verlag.

Detmar Meurers. 2009. On the automatic analysis of learner language: Introduction to the special issue. *CALICO Journal*, 26(3):469–473.

Jürgen Quetz. 2001. Der gemeinsame europäische referenzrahmen. *Info DaF*, 28(6):553–563.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas, 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen*. Humboldt Universität zu Berlin, 2.01 edition.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus: A flexible multi-layer corpus architecture. In Nicolas Ballier Díaz-Negrillo, Ana and Paul Thompson, editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. Amsterdam: John Benjamins.

Emanuel A Schegloff. 1988. Presequences and indirection: Applying speech act theory to ordinary conversation. *Journal of Pragmatics*, 12(1):55–62.

Emanuel A Schegloff. 1996. Issues of relevance for discourse analysis: Contingency in action, interaction and co-participant context. In *Computational and conversational discourse: Burning Issues – An Interdisciplinary Account*, pages 3–35. Springer-Verlag Berlin Heidelberg.

David Schlangen. 2004. Causes and Strategies for Requesting Clarification in Dialogue. In *5th Workshop of the ACL SIG on Discourse and Dialogue*.

Thomas Spranz-Fogasy. 2002. *Interaktionsprofile: Die Herausbildung individueller Handlungstypik in Gesprächen.* Radolfzell: Verlag für Gesprächsforschung.

Tanya Stivers and Nick J Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626.

Joel R Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 24–32. Association for Computational Linguistics.

Vincenza Tudini. 2010. *Online Second Language Acquisition: Conversation Analysis of Online Chat*. Continuum.

Heike Zinsmeister and Margit Breckle. 2012. The alesko learner corpus: design–annotation–quantitative analyses. *Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: John Benjamins*, pages 71–96.