ERROR ANNOTATION IN SPOKEN LEARNER CORPORA

Anke Lüdeling, Simon Sauer, Malte Belz, Christine Mooshammer

Humboldt-Universität zu Berlin
corresponding author: anke.luedeling@rz.hu-berlin.de

Keywords: multi-layer annotation, error annotation, spoken learner corpus, BeMaTaC

1. INTRODUCTION

In order to find learner-specific linguistic properties, patterns in learner corpora are often analyzed quantitatively and compared to patterns in other learner corpora or native speaker corpora (see e.g. [11, 12, 23, 13]). Some patterns can be found on the surface of the learner text using word forms or properties of the sound signal but many research questions require the analysis of more abstract patterns involving, for example, phonemes, tones, lemmas, parts of speech, syntactic phrases, or error categories. This paper is largely methodological and focuses on the question of how error annotation can be done consistently and transparently in a spoken learner corpus. We will illustrate our points with data from the *Berlin Map Task Corpus* (BeMaTaC v. 2013-02.1, [22], see Section 2).

There are at present not many spoken learner corpora. Only some of the existing spoken learner corpora contain sound files, and only some of these are time-aligned and stored in a multi-layer corpus architecture where different annotation layers can be added freely (see [2] for a typology of spoken learner corpora and a discussion of these issues). Some spoken learner corpora are produced for phonetic questions and annotated and analyzed in a tool dedicated for phonetic phenomena. Other spoken learner corpora are collected for lexical, syntactic, or communicative purposes and these often do not contain the signal. Annotation is done using tools dedicated for token, or span annotation, syntactic annotation, or sometimes pointing relations. We will argue below that many properties of learner language and learner speech can only be understood through the combination of information on many layers. This implies a corpus architecture that allows annotation through different tools that are then merged into a common corpus.

In Section 3 we argue that error identification implies the implicit or explicit formulation of a target hypothesis, and that there can be different target hypotheses for the same text depending on the research question and desired granularity. Since error annotation can pertain to phonetic phenomena in learner speech as well as to grammatical or even communicative properties of learner language - and all of these concurrently in the same corpus - we use a corpus architecture which allows for the alignment of the signal to a transcript, multiple tokenizations and as many annotation layers as necessary [15, 21].

2. BEMATAC

The Berlin Map Task Corpus (BeMaTaC; https://u.hu-berlin.de/bematac) is a freely available corpus of spoken German. It consists of an L1 subcorpus recorded with native speakers of German and an identically designed L2 subcorpus with advanced speakers of German as a foreign language (to date, all learners in the corpus are native speakers of English and have test scores equivalent to ECFR level C1 or above). BeMaTaC uses a map-task design, where one speaker (the instructor) instructs another speaker (the instructee) to reproduce a route on a map with landmarks [1]. The corpus is accessible via ANNIS [15], an open-source browser-based search and visualization tool.

3. ERROR ANNOTATION AND SPOKEN LEARNER CORPORA

3.1. Error identification

Error annotation is a difficult task (see [16] for a more thorough discussion). The main reason for this is conceptual: It is not always clear what constitutes an error. This has been discussed extensively in the literature on second language acquisition and foreign language teaching, and there are many suggestions for a definition of ‘error’, some involving purely grammatical criteria, others focusing more on the adequacy of an utterance in a given context, the comparison of what a learner does with what a native speaker would do in a given situation, etc. In essence, however, there can be no general definition of error, and the decision of what constitutes an error depends on the research goal (see, among many others, [6, 5, 9, 10, 7, 19]).

The first step in error annotation is error identification, i. e. a decision on the exponent of the error. Even if the research goal is clear and a precise error definition can be derived from it, it is often unclear how to interpret a learner utterance. Each error is a difference between the utterance and an explicit or implicit 'correct' utterance. This is sometimes called **target hypothesis**. Here we define 'error' as the difference between the learner utterance and a target hypothesis. There can be many target hypotheses for a given learner utterance. A target hypothesis does not constitute the 'truth' or the 'only correct way of saying something' but is an interpretation of the utterance for the purpose of a given research goal [17].

We want to illustrate this using a purely grammatical notion of error and two examples from a written learner corpus containing texts from advanced learners of German as a foreign language (the Falko corpus, [18]). (1) contains a number mismatch between an adjective and the noun it modifies. This can be 'corrected' in several ways: the number of the adjective can be changed, the number of the noun can be changed, or the noun phrase can be labeled as a whole. Each of the error marking strategies can be defended. In an error analysis the different strategies would lead to different error counts on adjectives, or nouns. The verb *erlernen* in (2) does not subcategorize for a reflexive and, while being possible, it is not the ideal verb here. One could 'correct' this sentence in several ways, and again the target hypothesis will influence the error analysis that follows: delete the reflexive (→ argument structure error), change the verb (for example to *aneignen* "to acquire" → lexical/stylistic error), or do both (*neues Wissen zu erwerben* → argument structure error and lexical error).

(1) Um die richtige Strategien in diesen
in-order-to the right.SG strategies.PL in these
Bereichen wählen zu können
areas to-choose to be-able
'In order to be able to choose the right strategie(s) in these areas'

(2) bevor man überhaupt anfangen kann, sich neues
before one even start can, REFL new
Wissen zu erlernen
knowledge to learn
'before one can even start to acquire new knowledge'

The consequence of these issues is that it is necessary to construct an **explicit** target hypothesis (or several) according to transparent criteria (see [20] for a description of several target hypothesis pertaining to different research questions in the Falko corpus). It is equally necessary to construct a target hypothesis following the same criteria for each corpus the learner corpus is compared to. A shared baseline is essential, as native speakers do not always behave in a way grammar would predict. Constructing target hypotheses is difficult even for fairly advanced, written learner language. It becomes more difficult for varieties that are further away from a 'standard'.

3.2. Schwa elision

We argued that the comparison of patterns in learner corpora and native speaker corpora across several annotation layers leads to interesting acquisition results. We want to briefly illustrate our point by looking at final schwa in German. In spontaneous German speech, schwa elision occurs quite frequently in word-final position (this is a reduced account; we are aware of the fact that schwa/non-schwa is not a binary decision and that many phonetic parameters have to be taken into account; for a thorough study see e.g. [14]). In BeMaTaC, we can find instances of final schwa elision through a comparison between the diplomatic (narrow transcription) and the normalized transcription, cf. Table 1.

Table 1: Example of the multi-layer architecture in BeMaTaC.

dipl	das	hab	ich	nicht	gesagt
norm	das	habe	ich	nicht	gesagt
pos	PDS	VAFIN	PPER	PTKNEG	VVPP
gloss	*that*	*have*	*I*	*not*	*said*

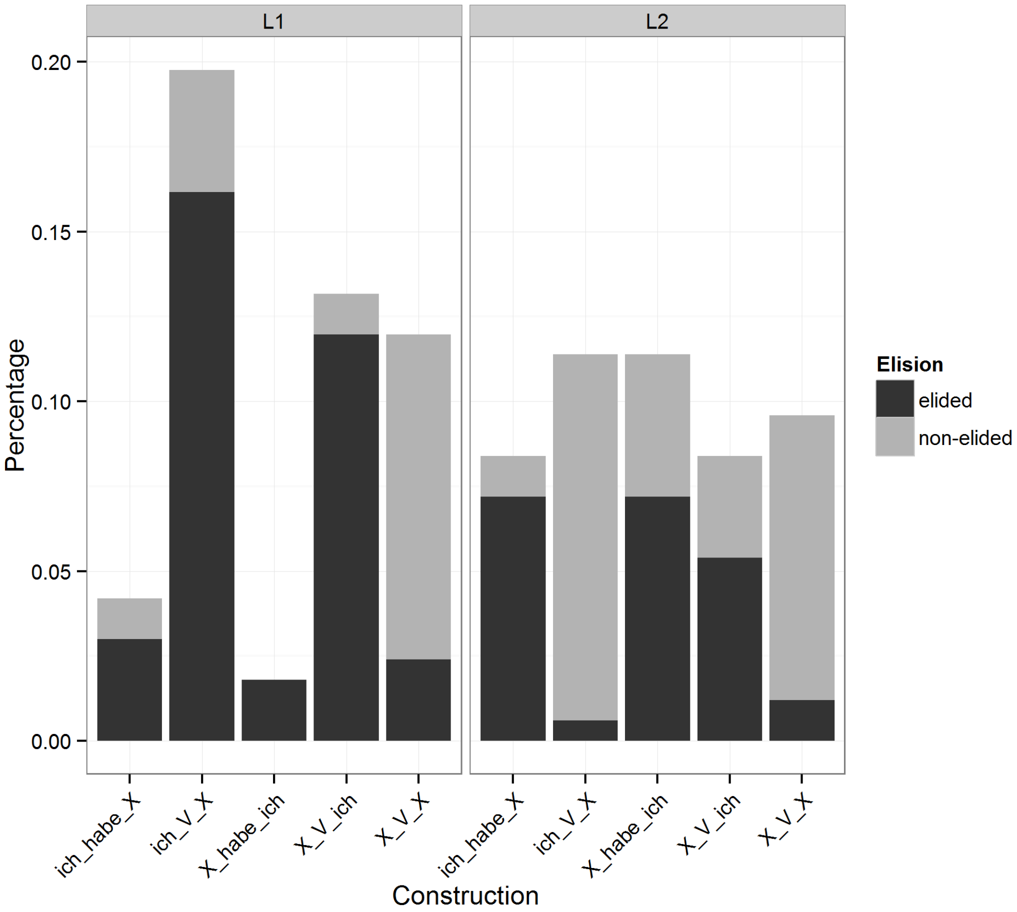
Here, the normalization can be used as a target hypothesis. We are able to integrate information from different linguistic annotations, such as part-of-speech tags or lemmatization. A first analysis reveals that both learners and native speakers do not elide schwas in nouns. Schwas in verbs, however, behave differently (imperatives are excluded from our analysis, as they are paradigmatically schwaless).

Table 2: Frequencies of schwa elision in BeMaTaC.

	∅-forms (dipl)	-*e*-forms (norm)	%
L1	67	108	62
L2	44	107	41

The interpretation of the patterns depends on the research agenda. When the normalized layer with unelided forms is seen as a target hypothesis, native speakers produce more 'errors' than learners (cf. Table 2). However, this target hypothesis reflects a conceptually written standard. When adhering to a setting of spontaneous speech, we

Figure 1: 3-gram constructions of finite verb forms ending with *-e* in BeMaTaC. V stands for all verbs other than *haben* 'have', X stands for any element except *ich* 'I'



may conclude that learners have not yet achieved the level of schwa elision that is typical of native speakers. A more detailed comparison reveals interesting patterns of 3-gram constructions (cf. Figure 1). The most prominent difference is that L1 speakers use schwa elisions more productively, with a wider range of verbs (12 hapax legomena, e.g., *beschreib* 'explain' or *find* 'find'). L2 speakers, on the other hand, predominantly elide schwas in the construction *ich_habe_X/X_habe_ich* 'I_have_X/X_have_I', which may serve as a teddy-bear construction [8] in the acquisition of verb-final schwa elision.

Spontaneous speech deviates from written language in many other ways. The comparison between a diplomatic layer and a normalized layer in usage data allows us to find these instances and compare overuse and underuse between native speakers and learners.

4. CONCLUSION

We have shown that the definition of 'error' depends on the underlying concept of a target hypothesis. The target hypothesis must be defined according to the research question. Therefore, multiple target hypotheses can be applied. Target hypotheses can include or even combine various linguistic domains, such as phonetics, morphology and syntax. This is only possible using a multi-layer corpus architecture.

5. REFERENCES

[1] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34, 351–366.

[2] Ballier, N., Martin, P. 2015. Speech annotation of learner corpora. In: Granger, S., Gilquin, G., Meunier, F., (eds), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

[3] Belz, M. 2013. Disfluencies und Reparaturen bei Muttersprachlern und Lernern – eine kontrastive Analyse. Master's thesis Humboldt-Universität zu Berlin.

[4] Belz, M., Sauer, S., Lüdeling, A., Mooshammer, C. 2015. Repair behaviour of advanced German learners in the Berlin Map Task Corpus. Trouvain, J., Zimmerer, Frank, Gásy, Mária, , Bonneau, A., (eds), *IFCASL Workshop on Phonetic Learner Corpora*.

[5] Bley-Vroman, R. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17.

[6] Corder, S. P. 1981. *Error Analysis and Interlanguage*. Oxford: Oxford University Press 4. impr. edition.

[7] Díaz-Negrillo, A., Fernández-Domínguez, J. 2006. Error tagging systems for learner corpora. *Revista española de lingüística aplicada / RESLA* 19(19), 83–102.

[8] Ellis, N. C. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32, 17–44.

[9] Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford University Press, USA.

[10] Ellis, R., Barkhuizen, G. 2009. *Analysing Learner Language*. Oxford applied linguistics. Oxford: Oxford Univ. Press [nachdr.] edition.

[11] Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In: Aijmer, K., (ed), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies Lund 4 - 5 March 1994* volume 88 of *Lund studies in English*. Lund: Lund University Press [u.a.] 37–51.

[12] Granger, S. 2002. A bird's-eye view of learner corpus research. In: Granger, S., (ed), *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching* volume 6 of *Language learning and language teaching*. Amsterdam: Benjamins 3–33.

[13] Granger, S. 2008. Learner corpora. In: Lüdeling, A., Kytö, M., (eds), *Corpus Linguistics. An International Handbook* volume 1. Berlin: Mouton de Gruyter 259–275.

[14] Kohler, K. J., Rodgers, J. 2001. Schwa deletion in German read and spontaneous speech. *AIPUK* 35, 97–123.

[15] Krause, T., Zeldes, A. 2014. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*.

[16] Lüdeling, A., Hirschmann, H. 2015. Error annotation. In: Granger, S., Gilquin, G., Meunier, F., (eds), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

[17] Lüdeling, A. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter, M., Grommes, P., (eds), *Fortgeschrittene Lernervarietäten*. Tübingen: Niemeyer 119–140.

[18] Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., Walter, M. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 45(2), 67–73.

[19] Ragheb, M., Dickinson, M. 2011. Avoiding the comparative fallacy in the annotation of learner corpora. Granena, G., Koeth, J., Lee-Ellis, S., Lukyanchenko, A., Botana, G. P., Rhoades, E., (eds), *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions* Somerville, MA. Cascadilla Proceedings Project 114–124.

[20] Reznicek, M., Lüdeling, A., Hirschmann, H. 2013. Competing target hypotheses in the falko corpus: A flexible multi-layer corpus architecture. In: Díaz-Negrillo, A., Ballier, N., Thompson, P., (eds), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins 101–124.

[21] Sauer, S., Lüdeling, A. to appear. Flexible multilayer spoken dialogue corpora. *International Journal of Corpus Linguistics. Special Issue on Spoken Corpora*.

[22] Sauer, S., Rasskazova, O. 2014. BeMaTaC – eine digitale multimodale Ressource für Sprach- und Dialogforschung. Workshop Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen. Berlin 2014. Poster. Digital Humanities Conference, Berlin.

[23] Tono, Y. 2004. Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In: Aston, G., Bernardini, S., Stewart, D., (eds), *Corpora and Language Learners* volume 17 of *Studies in corpus linguistics*. Amsterdam and Philadelphia: John Benjamins 45–66.