EAGLE: an Error-Annotated Corpus of Beginning Learner German

Adriane Boyd

Department of Linguistics
The Ohio State University
adriane@ling.osu.edu

Abstract

This paper describes the Error-Annotated German Learner Corpus (EAGLE), a corpus of beginning learner German with grammatical error annotation. The corpus contains online workbook and and hand-written essay data from learners in introductory German courses at The Ohio State University. We introduce an error typology developed for beginning learners of German that focuses on linguistic properties of lexical items present in the learner data and that has three main error categories for syntactic errors: selection, agreement, and word order. The corpus uses an error annotation format that extends the multi-layer standoff format proposed by Lüdeling et al. (2005) to include incremental target hypotheses for each error. In this format, each annotated error includes information about the location of tokens affected by the error, the error type, and the proposed target correction. The multi-layer standoff format allows us to annotate ambiguous errors with more than one possible target correction and to annotate the multiple, overlapping errors common in beginning learner productions.

1. Introduction

Corpora of learner language provide useful data for research in language acquisition and the development of natural language technology. Learner productions provide insight into the language acquisition process and annotated learner corpora allow researchers to easily search for particular phenomena. Annotated data is also useful for developing and customizing tools such as part-of-speech taggers and spell checkers for non-native speakers (cf. Granger, 2003; Meurers, 2009). To support research in these areas for learners of German, we have created the Error-Annotated German Learner Corpus (EAGLE), which is the first freely available error-annotated corpus for beginning learners of German.

2. Data

The learner language data in the EAGLE corpus consists of responses to course-related activities from students in the second and third courses of The Ohio State University's introductory German sequence. Two main types of data were collected: online workbook responses and final exam essays. The two types of data were chosen to include both typed and hand-written language produced with and without access to reference materials.

2.1. Online Workbook Data

The online workbook subcorpus contains data collected from the *Deutsch: Na Klar! Online Workbook, 4th Edition* (Briggs, 2003). Responses were collected from 50 learners (38 in the second course and 12 in the third course) during one quarter at The Ohio State University. The online workbook contains a wide variety of activities including translation exercises, cloze questions, build-a-sentence questions, etc. which the learners completed outside of class with access to reference materials. A translation exercise with sample learner responses is shown in Figure 1.

The online workbook responses range from answers to multiple choice questions to short essays. In order to focus on data suited for grammatical error annotation, the EAGLE corpus contains responses to only those activities where the

Translate into German:

    To whom do these articles of clothing belong?

Sample responses:

Wem hat diesen Kleidungsstücke
Wer gehört diese Kleidungen?
Wem gehören diese Kleidungsstücke?
Wem gehört diesem Kleidungs?
Wem gehört die Kleidungsstücke?
Wer gehören diese Kleidungsstucke?
Wem gehört diesem Kleidungstücke?
Wem gehörnen deser Kleidungsstüke zu?
Wem gehören diese Kleidungsstucke?
Wem gehören dieser Kleidungsstüke?
Wem gehören diesem Kleidungsstucke?
Wem gehört diese Artikel der Kleidung?

Figure 1: Sample Exercise from Online Workbook

learners are instructed to respond in complete sentences. In the activities where responses were automatically assessed by the online workbook, students often made multiple submissions until they reached the target answer. Each of these responses is stored separately in the corpus.

In total, there are 59,068 tokens in 6,986 responses to 412 activities. When duplicate responses to the same activity are removed (since many students arrived at the same target answer for a given activity), there are approximately 33,000 tokens in 3,500 responses containing a total of 7,500 sentences.

2.2. Essay Data

The essay subcorpus contains hand-written essays from 81 learners (43 in the second course and 38 in the third course) collected during a different quarter at Ohio State.[1] The

---

[1]Due to the anonymous data collection, it is not possible to determine whether any of the same learners appear in both the

Morphological	*gestudiert/studiert (*studied*)
	*machtete/machte (*made*)
	*wollst/willst (*want*)
Typographical	*iher/hier (*here*)
	*heißr/heißt (*is called*)
	*KOffer/Koffer (*suitcase*)
Capitalization	*wetter/Wetter (*weather*)
	*maria/Maria
	*hut/Hut (*hat*)

Figure 2: Example Non-Word Spelling Errors

learners could choose from several different topics and the essays were written as part of a timed exam without access to reference materials. The hand-written data was keyed in and the subcorpus contains 12,412 tokens in 81 essays with an average of 16 sentences per essay.

2.3. Preprocessing

The collected data was tokenized using Stefanie Dipper's German tokenizer (Dipper, 2008) and then anonymized to remove all potentially identifying personal names, streets, cities, and states. In order to maintain coherence in longer responses, each anonymized item receives a code such as "CITY-4" or "FIRSTNAME-13" that is used consistently throughout the corpus.

3. Error Annotation

The EAGLE error typology and annotation format focus on the annotation of grammatical errors present in the learner data. Before the grammatical error annotation begins, non-word spelling errors are corrected as described in section 3.1. Then, the grammatical error typology described in section 3.2 is applied using the multi-layer standoff format described in section 3.3.

Each sentence in the corpus is annotated independently without regard to context. If there is no context in which the sentence could be uttered, a series of one or more corrections are annotated that transform the ungrammatical sentence into a grammatical one. Each correction includes information about which tokens are affected by the error, the type of error, and the proposed target correction.

3.1. Non-Word Spelling Errors

Non-word spelling errors were identified and corrected to either a word with the smallest edit distance or to a literal translation in the case of English or other foreign words. These corrections build a small spelling error corpus with 1,697 tokens for 1,234 type non-word spelling errors.

A sample of the spelling errors shown in Table 2 illustrates a wide range of error types. The spelling errors identified in the EAGLE corpus have not yet been systemically analyzed. For a detailed analysis of spelling errors by non-native writers of German, see Rimrott (2005).

3.2. Error Typology

The error typology, which is informed by two previous classification schemes from Rogers (1984) and Juozulynas

workbook and essay subcorpora, but it is unlikely that the two learner groups overlap.

(1994), who respectively addressed errors by advanced and intermediate learners of German. The error typology includes five main types of errors: word form (errors within single words that are not non-word spelling errors), selection, agreement, word order, and punctuation. Examples of each type of error are shown in Figure 3.

The error types related to grammatical errors – selection, agreement, and word order – focus on linguistic properties of the lexical items present in the data and the relations between these items. Detailed error annotation schemes for these types are shown in Figures 4–6. Each type of error is subcategorized by grammatical features of the words, phrases, or topological fields (Höhle, 1986) affected by the error.

For most error types, the annotation proceeds bottom-up by considering the relations between lexical items present in the data. For instance, determiner-adjective-noun agreement is checked whenever a noun phrase with a determiner or adjective is found in a response; if a sentence does not contain any such noun phrases, there is no need to consider determiner-adjective-noun agreement. Exceptions to this are the word order errors that examine the positions of topological fields in a top-down fashion and the "Sentence" selection error, which also checks top-down for the presence of main clauses and finite verbs in each sentence.

3.3. Error Annotation Format

The EAGLE grammatical error annotation uses a multi-layer standoff format first proposed for learner error annotation by Lüdeling et al. (2005) for the FALKO corpus of advanced learner German (Siemen et al., 2006). This format is chosen in order to account for situations where a) errors span multiple words, b) learners make multiple overlapping errors in a single sentence, and c) errors are ambiguous. Standoff annotation allows multiple overlapping errors to be annotated easily and multiple layers allow for multiple target corrections to be specified in the case of ambiguities.

As in Lüdeling et al. (2005), each type of error encompasses three layers in the annotation: location, description, and target. The location layer identifies which words, phrases, or clauses are affected, the description layer specifies the particular type of error such as a subject-verb agreement error, and the target layer gives the target correction that corresponds to the error description. The target correction makes explicit the annotator's hypothesis about the learner's intended utterance and shows the correction for the specified error.

Example (1) shows a sentence with multiple errors that will be used in the following sections to illustrate the annotation format. It contains a noun phrase *dieses Hunden* 'this dog' where the determiner and the noun disagree in gender, case, and number and a verb complement *Wen* 'whom' in the wrong case. First, the agreement error will be considered. Figure 7 shows the appearance of the standoff annotation layers for agreement errors in this example.

(1) * Wen gehört dieses Hunden?
    whom$_A$ belong$_{3,sg}$ this$_{neut,N/A,sg}$ dog$_{masc,D,pl}$
    'Whom does this dogs belong?'

Error Type	Example	Detailed Error Description
Word Form	Ja, **Ich** zeige ihn ihnen. *yes, I show him them* Target: Ja, ich zeige ihn ihnen.	Capitalization
Selection	Hast du **der Reiseprospekt**$_{nom}$? *have you the travel brochure$_{nom}$* Target: Hast du den Reiseprospekt?	Verb - NP Complement Case
Agreement	**Du arbeiten** in Liechtenstein. *you work$_{1st/3rdplural}$ in Liechtenstein* Target: Du arbeitest in Liechtenstein.	Subject-Verb Agreement
Word Order	Welcher Job diese Dinge verlangen **würde**? *which job these things require would* Target: Welcher Job würde diese Dinge verlangen?	Finite Verb Position
Punctuation	Gehört dir diese Jacke *belongs you this jacket* Target: Gehört dir diese Jacke?	Missing Sentence-Final Punctuation

Figure 3: Types of Errors Annotated

Tokens	Wen	gehört	dieses	Hunden?
Location				
Description				
Target				

Figure 7: Agreement Error Annotation Layers

If we want to annotate the agreement error in *dieses Hunden*, we identify the affected tokens, determine the type of error, and give a target correction as in Figure 8 below.

Tokens	Wen	gehört	dieses	Hunden?
Location			1	
Description			Det-Noun Agreement	
Target			dieser Hund	

Figure 8: Agreement Error Annotation

3.3.1. Incremental Analysis

Because responses from beginning learners often contain multiple errors (cf. Heift, 2003), we extend the basic annotation format of Lüdeling et al. (2005) to include error numbering, which is specified in the location layer for each error. The error numbering allows the annotator to specify a series of incremental corrections, each with its own detailed error description, that convert the learner's response into a grammatical target. Each step assumes that previous corrections have been made, which allows us to address phrase-internal errors, such as agreement errors, before considering selection or word order. For example, all of the words in a noun phrase need to have the same number, gender, and case before it is possible to determine whether that noun phrase is grammatical as a particular complement of a verb.

In example (1) from the previous section, the subject *dieses Hunden* 'this dogs' needs to be internally consistent before an annotator can determine whether the subject agrees with the verb *gehört* 'belong'. In this case, the phrase *dieses Hunden* 'this dogs' would be annotated as containing a determiner-noun agreement error with the target cor-

rection *dieser Hund* 'this dog (nom sg)' and once this is complete, the subject-verb agreement can then be examined and be determined to be grammatical. After examining the subject-verb agreement, we can turn to the other verb complement from the example, *Wen* 'whom'. Instead of an accusative complement *Wen*, the verb *gehört* requires a dative complement *Wem* 'to whom'. The annotation including both the agreement error and this verb complement case error is shown in Figure 9 below.

Tokens	Wen	gehört	dieses	Hunden?
Agr. Loc.				1
Agr. Desc.			Det-Noun Agreement	
Agr. Target			dieser Hund	
Sel. Loc.	2			
Sel. Desc.	NP Compl. Case			
Sel. Target	Wem			

Figure 9: Incremental Error Annotation

When the two target corrections in Figure 9 are applied to the original sentence, we arrive at a grammatical target sentence:

(2) Wem gehört dieser Hund?
 to whom$_D$ belong$_{3,sg}$ this$_{masc,N,sg}$ dog$_{masc,N,sg}$
 'To whom does this dog belong?'

For example (1), the order in which in the errors are annotated is not important because they do not overlap, but in many instances, the order in which the errors are annotated plays an important role in making it possible to annotate errors that depend on previous target corrections. We will return to the issue of overlapping errors after discussing ambiguous errors in the next section.

3.3.2. Dealing with Ambiguous Errors

Example (1) also illustrates how ambiguous errors, such as a large percentage of agreement errors, can cause difficulties in creating consistent annotation. Considering only the

S1. Verb

  A. Complement

    i. NP complement - incorrect case

    ii. PP complement - incorrect preposition

    iii. PP complement - incorrect case with correct preposition

    iv. Two-way PP complement with verb of state/location - incorrect preposition or case

    v. Two-way PP complement with verb of motion - incorrect preposition or case

    vi. VP complement - haben/sein error

    vii. VP complement - incorrect non-finite verb form

    viii. Clausal complement - incorrect complementizer

    ix. Incorrect complement type

    x. Missing

    xi. Extra

  B. Separable prefix - impossible form

  C. Reflexive

    i. Missing

    ii. Extra

    iii. Incorrect case

S2. Preposition

  A. Complement

    i. Incorrect case

    ii. Missing

S3. Noun

  A. Determiner

    i. Missing

    ii. Extra

  B. Complement

    i. NP complement - incorrect case

    ii. PP complement - incorrect preposition

    iii. PP complement - incorrect case with correct preposition

S4. Adjective

  A. Complement

    i. NP complement - incorrect case

    ii. PP complement - incorrect preposition

    iii. PP complement - incorrect case with correct preposition

    iv. Incorrect complement type

  B. Comparative clause

S5. Sentence

  A. Main clause

    i. Missing

  B. Finite verb

    i. Missing

    ii. Extra

Figure 4: Selection Error Typology

A1. Subject-Verb

  A. Person

  B. Number

A2. Determiner-Adjective-Noun

  A. Gender

  B. Number

  C. Case

  D. Definiteness

  E. Attributeness

A3. Relative Pronoun-Antecedent

  A. Gender

  B. Number

  C. Case

A4. Subject-Predicate with Copula

  A. Number

A5. Reflexive-Subject

A6. Appositives

  A. Gender

  B. Number

  C. Case

Figure 5: Agreement Error Typology

O1. Finite verb

  A. In a main clause

  B. In a subordinate clause

O2. Non-finite verb

O3. Separable prefix

O4. Mittelfeld

  A. Arguments

  B. Adverbs

O5. Prepositional phrase

O6. Noun phrase

O7. Adverb phrase

Figure 6: Word Order Error Typology

noun phrase from the previous example, an annotator could have just as easily corrected *dieses Hunden* 'this dogs' to *diesen Hunden* 'these dogs (D pl)', which would have had both the incorrect number and case as the subject of the sentence. This would have led to further corrections to reach a grammatical target. In order to avoid these kinds of inconsistencies, an annotator chooses the target that minimizes the total number of errors annotated for the given sentence. Thus, instead of trying to minimize the edit distance between the learner response and the target correction, as in many existing error annotation schemes, the EAGLE annotation tries to minimize the total number of annotated errors.

In cases where the ambiguity is not resolved by the surrounding context, the multi-layer annotation allows for multiple targets to be specified. Because ambiguities most

Corrected Tokens	Als	Sofie	last	dem	Fahrplan,	sie	Reisepläne	machte.
Selection Loc					2			
Selection Desc			Verb - NP Complement Case					
Selection Target				den				
Agreement Loc		1						
Agreement Desc 1		Subject-Verb						
Agreement Target 1			las					
Agreement Desc 2								
Agreement Target 2								
Word Order Main Clause Loc				4				
Word Order Main Clause Desc			Finite Verb Position - Main Clause					
Word Order Main Clause Target		Als Sofie den Fahrplan las, machte sie Reisepläne.						
Word Order Sub. Clause Loc		3						
Word Order Sub. Clause Desc		Finite Verb Position - Subord. Clause						
Word Order Sub. Clause Target		Als Sofie den Fahrplan las,						

Figure 10: Multi-Layer Standoff Annotation for Example (3)

often arise in agreement errors, the EAGLE annotation scheme includes two additional layers in the agreement type for a second error description and a second error target. The additional layers are shown in the example in Figure 10, which is described in detail in the next section.

3.3.3. Overlapping Errors

A final issue common in learner language productions is overlapping errors. Since different types of errors are annotated in different layers, the multi-layer standoff format makes it simple to annotate such errors. Example (3) shows what the multi-layer standoff format looks like for a response with multiple overlapping errors. This example, which combines errors from several actual learner responses, contains four errors: 1) a subject-verb agreement error, 2) a noun phrase argument in the wrong case, 3) a word order error in the subordinate clause, and 4) a word order error in the main clause. The EAGLE multi-layer standoff annotation for example (3) is shown in Figure 10. In order to show overlapping word order error spans, the word order error layers have been divided into two sets of layers for the main and subordinate clauses.

(3) Als Sofie last   dem Fahrplan, sie
    when Sofie read$_{2nd,pl}$ the timetable$_D$ she
    Reisepläne machte.
    travel plans made

    'As Sofie read the timetable, she made travel plans.'

3.4. EAGLE Corpus Annotation

We are using the Partitur ('musical score') Editor from the EXMARaLDA (Extensible Markup Language for Discourse Annotation) Project (Schmidt, 2001) to perform the annotation and will distribute the EAGLE corpus in EXMARaLDA XML format. The annotation of the online workbook subcorpus by a single annotator is complete. The frequencies of the main error types are summarized in Figure 11 and the most frequent errors are shown in Figure 12.

Word Form	523
Selection	1,570
Agreement	927
Word Order	238
Punctuation	742

Figure 11: Errors in the Online Workbook Subcorpus

4. Conclusion and Future Work

The EAGLE corpus is the first corpus of freely available error-annotated data for beginning learners of German and we hope that the error annotation will be useful for research in the areas of language acquisition and intelligent computer-aided language learning. Future work includes the annotation of the essay subcorpus and annotation by additional annotators in order to evaluate the inter-annotator agreement for our error annotation scheme. On the basis of this corpus, we also plan to explore the automatic detection and diagnosis of word order errors for beginning learners of German.

References

Briggs, J. (2003). *Deutsch: Na klar! Online Workbook, 4th Edition*. McGraw-Hill.

Dipper, S. (2008). Tokenizer for German. http://www.linguistics.ruhr-uni-bochum.de/~dipper/tokenizer.html.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO* 20(3), 465–480.

Error Category	Error Description	Count
Agreement	Subject-Verb – Number	354
Selection	Verb – NP Complement Case	329
Agreement	Det-Adj-Noun – Gender	255
Selection	Sentence – Finite Verb Missing	201
Selection	Preposition – Complement Case	197
Agreement	Subject-Verb – Person	154
Agreement	Det-Adj-Noun – Case	126
Selection	Verb – Complement Missing	121
Agreement	Det-Adj-Noun – Number	115
Selection	Verb – Two-Way PP Complement with Verb of State/Location	112
Word Order	Finite Verb – Main Clause	108

Figure 12: Most Frequent Grammatical Errors in the Online Workbook Subcorpus

Heift, T. (2003). Multiple Learner Errors and Meaningful Feedback: A Challenge for ICALL Systems. *CALICO* 20(3), 533–549.

Höhle, T. (1986). Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*. Göttingen, Germany.

Juozulynas, V. (1994). Errors in the Compositions of Second-Year German Students: An Empirical Study for Parser-Based ICALI. *CALICO* 12(1), 5–17.

Lüdeling, A., M. Walter, E. Kroymann & P. Adolphs (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*. Birmingham.

Meurers, D. (2009). On the Automatic Analysis of Learner Language. Introduction to the Special Issue. *CALICO Journal* 26(3), 469–473.

Rimrott, A. (2005). Spell Checking in Computer-Assisted Language Learning: A Study of Misspellings by Nonnative Writers of German. Master's thesis, Simon Fraser University.

Rogers, M. (1984). On Major Types of Written Error in Advanced Students of German. *International Review of Applied Linguistics in Language Teaching* XXII(1).

Schmidt, T. (2001). The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. In *Proceedings of the IRCS Workshop On Linguistic Databases, 11-13 December 2001*. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania.

Siemen, P., A. Lüdeling & F. H. Müller (2006). FALKO - ein fehlerannotiertes Lernerkorpus des Deutschen. In *Proceedings of Konvens*. Konstanz.