# Error annotation systems

**Chapter** · January 2015

DOI: 10.1017/CBO9781139649414.007



**2 authors**, including:

Hagen Hirschmann
Humboldt-Universität zu Berlin
**17** PUBLICATIONS **75** CITATIONS

SEE PROFILE

# 7

# Error annotation systems

Anke Lüdeling and Hagen Hirschmann

## 1 Introduction

The categorisation and investigation of errors made by foreign or second language learners is an interesting and fruitful way of studying accuracy and other aspects of learner language (Corder 1967, 1981; Dagut and Laufer 1982; Ellis and Barkhuizen 2005, among many others). In addition to being an analytical tool for assessing the 'quality' of a text, error analysis, if done correctly, sheds light on the hypotheses a learner has about the language to be learned. Missing or incorrect articles, for instance, can point us to a better understanding of the learner's ideas of definiteness; certain lexical errors tell us that a learner might not be able to use the appropriate register, etc. Error analysis (henceforth EA) is a research method and, as for any other method, there are a number of issues to take into account when applying it. These issues include the categorisation and assignment of error types as well as the (linguistic and extra-linguistic) contextualisation of errors. It is, for example, often necessary to consider the larger context in order to decide whether a definite article is required. Knowing the first language (L1) of a learner and the circumstances under which a text was produced can be crucial in understanding a register error. Since in the early days of EA, some of these methodological issues have sometimes been neglected, and EA has often been criticised (for an overview of the criticism, see e.g. Dagneaux et al. 1998).

It has also long been recognised that the study of language acquisition processes needs to be reproducible and testable. Complementing experimental data of various types, learner corpora can be a valuable source of data for reproducible studies of language acquisition – but only if they are well designed, well described and publicly available. Corpus data must be interpreted and categorised to be useful. In this chapter, we focus on one way of interpreting learner corpus data, namely error annotation as the explicit and transparent way of marking errors in a learner corpus. Error

annotation is one step in computer-aided error analysis (CEA), a term introduced by Dagneaux et al. (1998) to refer to error analysis conducted on the basis of learner corpora. We will describe how error annotation schemes are designed, how they can be queried, and which opportunities and problems error annotation brings.

## 2 Core issues

### 2.1 Annotation

Unannotated corpus data can be used for many research questions. But whenever one wants to search for categories of something – all finite verbs, all sentences under five words, all orthography errors – and not strings, it is useful to assign these categories to the corpus data. The utterance in example (1), for instance, is annotated with part-of-speech categories,[1] lemmas, and noun phrases. It is now possible to search for noun phrases that contain conjunctions or for noun phrases that use singular nouns without a preceding article.

(1)

| | *The* | *learner* | *requires* | *support* | *and* | *guidance* |
|---|---|---|---|---|---|---|
| part of speech | AT0 | NN1 | VVZ | NN1 | CJC | NN1 |
| lemma | the | learner | require | support | and | guidance |
| noun phrases | NP | | | NP | | |

Error annotation works in the same way; segments from a learner corpus are annotated with an error category. Technically, annotation is the assignment of a category to a segment of the corpus (see also Chapter 5, this volume). It is done in the corpus and not somewhere else such as on a file card, in a spreadsheet, or in a statistical table.[2] Often the category is taken from a finite tagset, as in the part-of-speech layer in example (1) or, as we will see, an error category from a predefined error tagset. Sometimes this is not possible because the values that can be used are infinite or unforeseeable, as in a lemma layer or in the target hypothesis layers that we will introduce below.

Annotation is categorisation and thus involves a necessary loss of information. The same data can be categorised in different ways, even for the same type of information, depending on the criteria one wants to use. There are, for example, many part-of-speech tagsets (see Atwell 2008), some focusing on

[1] The utterance is taken from the *British National Corpus*, lemmas and noun phrases are added by us. The part-of speech tags are from the *CLAWS* tagset (Garside and Smith 1997); AT0 stands for article, NN1 for singular noun, VVZ for finite verb, CJC for conjunction.

[2] Linguistic data itself can be spoken or written. In this chapter we assume that the sound waves constituting spoken language data are represented by some kind of written representation, be it an IPA transcription or an orthographic transliteration, which then will be the base for further annotation as discussed here (see e.g. Lehmann 2004; Himmelmann 2012; Chapter 6, this volume).

the syntactic properties of words, others on the morphological properties, etc. An annotation layer thus never codes the 'truth' – rather it codes one way of interpreting the corpus data. Explicitly annotating the data means that the interpretation of the data is available to the reader of the analysis. It is important that annotation is separable from and will not corrupt the corpus data (Leech 2005). The corpus is separated into tokens – tokens are technically just the smallest units in a corpus and could be of any length and complexity – but typically in European languages they constitute something like 'graphemic words' – with all the problems that notion entails (see, e.g., Schmid (2008) for a discussion). Annotation can pertain to any unit of the corpus. The sequence in the corpus to which a category applies is called an exponent. There is 'subtoken' annotation such as phonetic or phonological annotation, token annotation such as part-of-speech annotation, annotation spanning several tokens, such as the annotation of the noun phrases in (1), idiomatic sequences, sentences, or paragraphs. The annotation itself can come in various formats. Next to the assignment of a simple category to a given token (such as a part-of-speech category 'NN1' to a singular noun) or a sequence of tokens, we find different types of hierarchical annotation (such as constituency trees), or pointing relations (such as the members of an anaphoric chain).

## 2.2 Corpus architectures

Because error annotation can, in principle, occur in different formats and attach to any exponent, and because there can be many layers of error annotation (see Section 2.5.1), a multi-layer corpus standoff architecture is very useful. In standoff architectures (see Carletta et al. 2003; Chiarcos et al. 2008, among many others) it is possible to define as many independent annotation layers as necessary. This means that different people, using different tools, can work on the same data and all their analyses can be consolidated. It also means that different interpretations of the same data can be kept apart. In accordance with what we said at the beginning, this makes it possible to test different hypotheses on the same data. We will show below that computer-aided error analysis often uses other annotation layers such as parts of speech or syntactic annotation in addition to the error categories, which is another reason why multi-layer architectures are helpful. That said, many existing learner corpora are not coded in standoff corpus architectures but use some kind of inline format where the annotation is not represented separately from the corpus data.

## 2.3 Error analysis

Errors may concern language production as well as language reception. In the following, we will exclusively discuss language-production errors.

Since the 1960s, the notion of what constitutes an error, how errors can be classified, and what role errors play in language acquisition has changed. We do not have space here to give an overview of the history of EA but can only sketch some of the major trends. For more comprehensive overviews of error analysis, see Corder (1981: 35ff.), Ellis (1994: 47ff.), James (1998) and Díaz-Negrillo and Fernández-Domínguez (2006).

The scientific study of learner errors is based on the assumption that errors are a surface reflex of the learner's internal grammar, or interlanguage (Selinker 1972). This notion was influenced by generative models of grammar that assume a systematic internal grammar for native speakers. The interlanguage of a learner is assumed to be just as systematic, although different from the internal grammar of native speakers.

Influenced by the idea of a systematic interlanguage, the perception of errors as expressions that are simply ill-formed and chaotic has given way to a concept of describing errors in a systematic way. Single errors are not useful for the study of language acquisition because an error might occur for any number of reasons and only some of these reasons might have to do with the learner's interlanguage, others being 'performance' errors due to tiredness, inattention, etc. It is thus necessary to study types of errors with many tokens and compare contexts and situations. One distinction that mirrors this is the error vs mistake distinction. 'Performance' errors – called mistakes – might point to processing issues but are not relevant for the study of interlanguage. 'Competence' errors (or simply 'errors'), on the other hand, might point to non-target-like structures in the interlanguage. While this distinction might be theoretically valid, it cannot be made in a corpus analysis because there is typically no way of knowing what the learner knew and which other tasks, feelings, etc. might have influenced his or her production. As an example, consider a (simplified) typing issue. The layout of a keyboard influences the frequency of typing mistakes. Keys that are next to each other are substituted for each other more often than keys that are further away from each other. On a qwerty-keyboard this could explain a number of *n* for *m* substitutions and it might be argued that they are simply mistakes and not influenced by the learner's knowledge of the target grammar. What happens, however, if a learner of L2 German substitutes *n* for *m* at the end of a word that should be in dative case (the accusative article *den* instead of the dative article *dem*, say)? This could be a pointer to an interlanguage problem or it could just be a typo. This simple example shows that the distinction between errors and mistakes can be made *after* a careful analysis of the data but not in the error annotation itself.

In tandem with the idea that errors can be described systematically, the notion of the function of errors in the acquisition process also changed (cf. Corder 1967, 1981). Errors used to be interpreted as violations of language rules that should be avoided. Today, certain types of errors are seen as signalling necessary stages on the way to target-like language. Many

of these necessary errors have to do with segmentation and productivity. Roughly speaking, learners first learn complex forms, such as inflected or derived words, without segmenting them. In that stage, they might not make errors but cannot use the language productively. At later stages, they might understand how to segment some of the forms and detect regularities, which leads to an overgeneralisation because they have not yet learned the exceptions. They might actually make more errors at this stage but these errors are a sign of a deeper understanding of the language and are therefore considered to be crucial in the acquisition process.

The idea that certain errors are typical for a given acquisition stage is fundamental in some acquisition models. Klein and Perdue (1997), for instance, hypothesise a common language-independent stage for untutored adult second language learners, which they call the basic variety. After observing second language learner groups with ten distinct first-language–second language-relations, Klein and Perdue describe common structural properties that occur for each group independent of the L2 and the L1. One example of such a property is 'no inflection in the basic variety, hence no marking of case, number, gender, tense, aspect, agreement by morphology' (Klein and Perdue 1997: 11). These properties are said to appear at a certain stage in the acquisition and will (for many learners) be overcome at some point when they proceed in their acquisition.[3] In a model like this, the appearance and disappearance of certain types of errors can be treated as benchmarks for the acquisition process.

The idea that a learner's interlanguage is systematic and that it can be analysed by looking at errors has been criticised repeatedly. One of the most fundamental pieces of criticism stems from Bley-Vroman (1983).[4] He states that error analysis is always done from a native-speaker perspective and that the analysis of a learner variety through the native perspective will not reveal the true properties of the learner's text. Bley-Vroman calls this the 'comparative fallacy'; similar issues have been raised by other researchers (Klein and Perdue 1997, for instance, use the term 'closeness fallacy' for a similar problem). Bley-Vroman (1983: 2) goes as far as to say that the comparative fallacy pertains 'to any study in which errors are tabulated … or to any system of classification of interlanguage (IL) production based on such notions as omission, substitution or the like'. Bley-Vroman's criticism is valid and any error annotation that uses a native 'standard' against which an error analysis is performed is problematic. This certainly has to be taken into account and there are many researchers that try to do error analysis in ways that avoid (or minimise) the comparative fallacy, mainly by carefully explaining and motivating any step in the analysis so as to find possible biases (see Tenfjord, Hagen and Johansen 2006; Ragheb and Dickinson 2011; Reznicek et al. 2013).

[3] This is a simplified account of Klein and Perdue's acquisition model. Other influences on learner language development discussed in that model are information structure and communicative needs.

[4] Bley-Vroman doubts the systematicity of interlanguage itself but this does not have to concern us here.

Ellis (1994: 50ff.) names four distinct steps in error analysis: (1) identification, (2) description, (3) explanation, and (4) evaluation of errors. It is important to keep these steps apart – conceptually and technically.

## 2.4 Identification and description of errors

In order to distinguish errors from non-errors in learner utterances it would be helpful to have a clear definition of what an error is. The first idea that comes to mind is that an error is a violation of a rule. To operationalise this, everything about our linguistic behaviour would have to be codified and described by rules. This is, of course, not possible. Many linguistic models make a distinction between grammar and usage, or between grammar errors and appropriateness errors.[5] Grammar, it is argued, can be codified by (categorial) rules, and grammar errors are therefore easy to detect and describe. Usage, on the other hand, is quantitative rather than categorial and appropriateness errors depend more on interpretation. If this were true, people would always agree on grammatical errors but we would expect some disagreement over the identification and classification of appropriateness errors.

Lennon (1991: 182) seems to refer to both types of error when he defines an error as 'a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native counterparts'. This definition results from the observation that 'to be fully nativelike language must be not only grammatical but also appropriate' (Lennon 1991: 184). Unidiomatic expressions or stylistically inappropriate forms can be grammatical in a strict sense. In Section 2.5.1 we will outline a way to deal with the difference between grammar and appropriateness errors.

In the following, we will argue that the identification and classification of grammar errors is far from easy and that already here we have to interpret the data. In order to address how a given learner expression would be used by native speakers it is necessary to provide an alternative expression. Compare examples (2)–(4).

(2) She must saves money.
(3) She must saved money.
(4) She must my.

Example (2) is clearly ungrammatical. We could provide a grammatical equivalent by changing the verb form: *She must save money*. This assumption makes it plausible to interpret the error as a verb morphology error

[5] The term 'usage' here is very broad and encompasses phenomena people have described as belonging to pragmatics, information structure, register, etc.

(no 3rd person-*s* in the scope of a modal) or predicate structure error. The assumption of which 'correct' utterance corresponds to the erroneous utterance is called reconstruction, target form or target hypothesis. For example (3), we can construct several target hypotheses which seem equally likely. Without further context, the following are equally possible and plausible: *She must have saved money* or *She must save money*. The error would be analysed differently depending on which target hypothesis is chosen. If the context does not help to disambiguate between these possibilities it is impossible to say which target hypothesis is to be selected. Example (4) shows an utterance which does not provide enough information to formulate a target hypothesis. In order to create a corresponding grammatical sentence, one would have to add so much information that innumerable target hypotheses are possible. This means that, in (4), it is impossible to appropriately analyse the error that causes the ungrammaticality of the utterance. The only sensible error category would be 'uninterpretable'.

Examples (2)–(4) show that clear grammatical errors already involve interpretation. This is even more true for appropriateness errors.[6] Appropriateness is judged differently by different people and can involve all linguistic levels. Words can be inappropriate (e.g. the use of *maybe* instead of *perhaps* in an academic register), syntactic structures can be inappropriate (e.g. a sentence with three complicated sub-clauses might be inappropriate in a conversation with a child), etc. While it is sometimes possible to mark a grammar error by looking at a sentence in isolation, the identification of appropriateness errors needs linguistic and extra-linguistic context. But just like grammar errors, appropriateness errors can only be found with the help of a target hypothesis. The target hypothesis might look different from one that is constructed for clear grammar errors. One possible way of dealing with this problem is the introduction of several target hypotheses; another might be the use of task-based corpora where the purpose of a learner utterance is clearly constrained by the context.

Errors cannot be found and analysed without an implicit or explicit target hypothesis – it is impossible *not* to interpret the data. It is important to note that the construction of a target hypothesis makes no assumptions about what a learner wanted to say or should have said. The analyser cannot know the intentions of the learner. The 'correct' version against which a learner utterance is evaluated is simply a necessary methodological step in identifying an error.

[6] The distinction between grammaticality and appropriateness is similar (although not equivalent) to what Ellis (1994: 701) and others have called overt vs covert errors. Overt errors are 'apparent in the surface form of the utterance' while covert errors are visible only in a broader context, 'when the learner's meaning intention is taken into account'.

In some learner corpora, target hypotheses are not given explicitly. Considering the high cost of the analysis, this is understandable. It is, however, a problematic decision: an error-annotated corpus which does not provide target hypotheses hides an essential step of the analysis – this could lead to mistakenly assuming that the error annotation which is present in a corpus is the 'truth' or 'correct analysis' instead of just one among many interpretations (similar to what Rissanen 1989 calls 'God's truth fallacy'). This is why increasingly more learner corpora are offering explicit target hypotheses along with error classes.

In error annotation it is necessary to assign one or more categories to each error; sometimes a token contains several errors, such as a morphological error and an orthographic error – in such cases it should be possible to assign both. Errors can be categorised according to many criteria – the exponent (one token, multiple tokens, etc.), the grammatical level (syntax, morphology, register, etc.), and many more. Which categorisation scheme and level of granularity is chosen depends on the research purpose (see Section 2.5.2).

## 2.5 Error marking

In Section 2.3 and 2.4 we have argued that error analysis is an interpretation of the primary data on many levels:

1. The identification of an error always depends on a target hypothesis. There can be more than one target hypothesis for any given learner utterance.
2. Even with the same target hypothesis there can be many different descriptions of an error. Error categories depend on the research question, the grammatical model, etc.
3. There can be several explanations for each description.

Error annotation schemes differ widely with respect to what counts as an error, the format of error coding, scope, depth of analysis, etc. This is, of course, mainly due to the different research questions (which lead to different categorisations). In the following we will compare some of the more common strategies that are found in existing error annotation schemes. Rather than aiming for a comprehensive list of existing systems we will concentrate on the underlying conceptual issues.

### 2.5.1 Target hypotheses

As stated in Section 2.4, the identification and categorisation of errors depend on an implicit or explicit target hypothesis. In this section we want to show how target hypotheses can be made explicit and what can be done if there are several competing target hypotheses. This pertains to the error exponent (sometimes also called extent of an error, or error domain) as well as to the error category (sometimes called error tag).

Clear-cut grammar errors seem easiest to deal with. Consider example (5) from the *Falko* corpus.[7]

(5) dass eine Frau zu Hause bleibt, um sich um den Kindern und dem Haus zu kümmern [fk008_2006_07_L2v2.4]
'that a woman stays at home in order to care for the children$_{DAT}$ and the house$_{DAT}$'

The coordinated noun phrase *den Kindern und dem Haus* is in dative case while the verb *kümmern* subcategorises for an *um*-PP, which itself subcategorises for accusative case. In this sentence, the description of the error seems clear – we would analyse it as a case error or a subcategorisation error. However, what is the error exponent? Is it the coordinated noun phrase *den Kindern und dem Haus*? Or do we assume two errors – one for each of the conjuncts of the noun phrase? Our decision involves assumptions about the syntactic structure of the sentence with regard to case marking and coordination. Are these the same assumptions that we would make in example (6), which is analogous in many respects (we are here only interested in the sequence *dem Führerschein und das Fahrenlernen* and ignore the other problems in the sentence)? We have a verb that subcategorises for a certain PP (*mit* and dative) and a subcategorised coordinated noun phrase. The learner uses the correct form *dem Führerschein* but the incorrect *das Fahrenlernen* within one coordinative structure. Here the error exponent cannot be the coordinated noun phrase and it is much less clear what type of error we are seeing.

(6) […] kann auch mit dem Führerschein und das Fahrenlernen eines PKS verglichen werden. [cbs003_2006_09_L2v2.4]
'… could be compared to a driver's license$_{DAT}$ and learning$_{NOM/ACC}$ to drive a car.'

Whichever way we decide, it becomes clear that if the analysis is not made explicit there is always the danger that such problems are (inadvertently) overlooked and that parallel cases in the data are treated differently. This has consequences for the error count and the final analysis (see Lüdeling 2008 for an experiment).

Often it is impossible to give one clear target hypothesis. Consider example (7) (taken from Weinberger 2002: 30). Here we see a number mismatch between the subject and the verb which should be congruent. In a target hypothesis either the subject or the verb could be changed,

[7] The *Falko* (*Fehlerannotiertes Lernerkorpus*) corpus contains essays written by advanced learners of German as a foreign language (Lüdeling et al. 2008).

see (8). Depending on which target hypothesis we chose we might have a subject number error or a verb number error.

(7) Jeder werden davon profitieren.
'Each$_{SINGULAR}$ will$_{PLURAL}$ profit from this.'

(8)

| | | | | |
|---|---|---|---|---|
| LU | *Jeder* | *werden* | *davon* | *profitieren* |
| TH 1 | *Jeder* | *wird* | *davon* | *profitieren* |
| | 'everyone$_{SG}$' | 'will$_{SG}$' | | |
| TH 2 | *alle* | *werden* | *davon* | *profitieren* |
| | 'everyone$_{PL}$' | 'will$_{PL}$' | | |

In this and the examples that follow LU=learner utterance, TH= target hypothesis

There are several possibilities in cases like (7). One possibility is to look at the context and decide whether there is a cue in it that points to one or the other option. This might often be feasible but there are two problems. First, the option that the annotator chooses might be influenced by his or her research interest; he or she might even see only one of the options. Second, if analogous cases are sometimes resolved one way and sometimes resolved in a different way, it is impossible to do a systematic search. Better alternatives for handling such cases are to either consistently resolve them in the same way (say, always change the subject, independent of context) or to give them an abstract mismatch tag (here: subject–verb agreement mismatch).

While grammatical (or overt) errors may be difficult to analyse, appropriateness errors are even more challenging. Consider example (9) from the *International Corpus of Learner English, ICLE* (Dagneaux et al. 2005).

(9) It sleeps inside everyone from the start of being, it just waits for opportunity to arose and manifest itself.

(ICLE-CZ-PRAG-0018.2)

Here the learner is writing about negative traits like greed, stating that they are present in every human being and emerge in situations which support them. The utterance *It sleeps inside everyone from the start of being* is not ungrammatical in a strict sense but it is still not quite idiomatic. The sequence *from the start of being* is not likely to be used by 'the speakers' native counterparts' (Lennon 1991: 182). Expressions like *since birth* or *from the beginning* sound more native-like in this situation. The first decision is *whether* to mark this as an error – it is easy to see that 'idiomaticity' is gradual (Lennon's phrase 'in all likelihood' reflects that). Example (10) shows three possible target hypotheses for the first part of (9). Again, the error exponent (as well as the resulting error description) differs but these target hypotheses also differ in abstractness. Target Hypotheses 1 and 2 provide an alternative wording, while Target Hypothesis 3 is more

abstract and says only that this part of the learner utterance is unidiomatic, conflating an implicit target hypothesis with an error tag (the annotator is only able to know that this expression is unidiomatic if he or she knows a more idiomatic expression).

| (10) | LU | *it* | *sleeps* | *inside* | *everyone* | *from* | *the* | *start* | *of* | *being* |
|---|---|---|---|---|---|---|---|---|---|---|
| | TH 1 | *it* | *sleeps* | *inside* | *everyone* | *since birth* | | | | |
| | TH 2 | *it* | *sleeps* | *inside* | *everyone* | *from* | *the* | *beginning* | | |
| | TH 3 | *it* | *sleeps* | *inside* | *everyone* | UNIDIOMATIC | | | | |

Different target hypotheses are not equivalent; a target hypothesis directly influences the following analysis. The *Falko* corpus consistently has two target hypotheses – the first one deals with clear grammatical errors and the second one also corrects stylistic problems.

The need for such an approach becomes clear in (11).

(11) Dependance on gambling is something like dependance on drugs (…)

(ICLE-CZ-PRAG-0013.3)

The learner utterance in (11) contains a spelling error. The two occurrences of *dependance* have to be replaced by *dependence*. From a more abstract perspective, the whole phrase *Dependence on gambling* sounds unidiomatic if we take into account that the learner wants to refer to a specific kind of addiction. Similarly, *dependence on drugs* appears to be a marked expression as opposed to *drug addiction*. An annotation that wants to take this into consideration has to separate the description into the annotation of the spelling error and the annotation of the stylistic error in order not to lose one of the pieces of information. Example (12) illustrates this.

| (12) | LU | Dependance | on | gambling |
|---|---|---|---|---|
| | TH 1 | Dependence | on | gambling |
| | TH 2 | Gambling addiction | | |

The examples in this section show how important the step of formulating a target hypothesis is – the subsequent error classification critically depends on this first step. In order to operationalise the first step of the error annotation, one can give guidelines for the formulation of target hypotheses, in addition to the guidelines for assigning error tags, which also need to be evaluated with regard to consistency (see Section 2.6).

The problem of unclear error identification has been discussed since the beginning of EA. Milton and Chowdhury (1994) have already suggested that sometimes multiple analyses should be coded in a learner corpus. If

the target hypothesis is left implicit or there is only one error analysis, the user is given an error annotation without knowing against which form the utterance was evaluated. In early corpora (pre-multi-layer, pre-XML) it was technically impossible to show the error exponent because errors could only be marked on one token. In corpora that use an XML format it is possible to mark spans, and target hypotheses are sometimes given in the XML mark-up. Only in standoff architectures, however, is it possible to give several competing target hypotheses. Examples of learner corpora with consistent and well-documented (multiple) target hypotheses are the *Falko* corpus, the trilingual *MERLIN* corpus (Wisniewski et al. 2013) or the *Czech as a Second Language* corpus (Rosen et al. 2014).

### 2.5.2 Error tagsets

Error annotation systems[8] assign categories to errors. As stated above, the types and granularity of the error categories depend on the research question. Technically, there is an error every time the learner utterance differs from the target hypothesis. The error tag describes the type of the error within a given error annotation scheme. There are systems that annotate errors on all grammatical levels, and there are systems that tag only one specific type of phenomenon such as, for example, errors pertaining to the marking of modality or tense. Some error annotation systems assign grammatical, lexical or other linguistic error categories.

Consider example (13), which covers the second half of the sentence in (9), and example (14). Both examples contain target hypotheses which provide grammatical structures for the respective ungrammaticalities that can be found in the original learner utterances. The errors in the learner utterances are made visible by the deviations of the target hypotheses from the target forms. The analysis of these errors is provided in two different ways. The first we call edit-distance-based error tagging, a form-based description of the edit operations that have to be performed in order to generate the target form out of the original learner form. The second error tag is a linguistic interpretation of the deviation. Edit-distance-based annotation schemes consist of categories like 'change', 'delete', 'insert', sometimes 'move-source', 'move-target' or what Lennon (1991: 189) calls errors of substitution, over-suppliance, omission or permutation. Once a target hypothesis is given, this can be done automatically. While a distance-based annotation scheme might not look very interesting in itself, it can become very useful in combination with other layers of annotation, such as part of speech or lemma (for an example, see Reznicek et al. 2013). One could then find all cases where an article was inserted or deleted.[9] Distance-based systems are often only the first step in the analysis – it is possible to add linguistically motivated

[8] They are sometimes called error taxonomies, a term we avoid because many of them are not taxonomies in a technical sense.

[9] There is another equally likely target hypothesis where *an* is changed into *some*. This would yield different error tags.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (13) | LU | *It* | *just* | *waits* | *for* | | *opportunity* | *to* | *Arose* | | *and* | *manifest* | *itself* |
| | TH | *It* | *just* | *waits* | *for* | DET | *opportunity* | *to* | *Arise* | | *and* | *manifest* | *itself* |
| | D-BT | | | | | INSERT | | | CHANGE | | | | |
| | L-BT | | | | | missing article | | | inflection // orthography | | | | |

D-BT = distance-based tagging, L-BT = linguistically based tagging

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (14) | LU | *He* | *tries* | *to* | *get* | *an* | *information* | *about* | *this* | *profession* |
| | TH | *He* | *tries* | *to* | *get* | | *information* | *about* | *this* | *profession* |
| | D-BT | | | | | DELETE | | | | |
| | L-BT | | | | | superfluous article | | | | |

error types on further annotation layers. Linguistically based tagging systems interpret the difference between the learner utterance and the target hypothesis with respect to a given grammatical or pragmatic model. For the *arose* case in example (13) they could use a tag for orthographic errors, for inflectional errors, or use an ambiguous or 'undecided' tag.

Díaz-Negrillo and Fernández-Domínguez (2006) discuss different linguistically motivated error-tagging systems for learner corpora, pointing out that 'the way the linguistic information is organised in taxonomies varies from system to system' (2006: 93). Often the error tags are conflated with part-of-speech or word-class information. As just one example, consider how sentences (13) and (14) would be tagged according to the *ICLE* 'Error Tagging Manual' (Version 1.2; Dagneaux et al. 2005). Errors there are divided into eight major categories: 'form'; 'grammar'; 'lexico-grammar'; 'lexis'; 'word redundant', 'word missing', 'word order'; 'punctuation'; 'style'; 'infelicities'. In 'grammar', 'lexico-grammar' and 'lexis', word classes are referred to explicitly, which means that the interpretation of many learner errors directly depends on the part of speech involved. Other errors depend on morphological, graphemic, word placement or stylistic problems, regardless of a specific word class (e.g. 'word order'). For missing elements like the article in (13) or superfluous elements like the article in (14) there is no intuitive method to anticipate whether to assign the error to the category word class (article) or to the fact that they are missing, but the manual tries to provide an unambiguous solution for all errors. According to Dagneaux et al. (2005), the missing article in (13) receives the error tag 'GA' for the classes 'grammar-article', the superfluous article in (14) receives the tag 'XNUC' (for the classes 'lexico-grammar-noun, uncountable/countable'), and the error in the form *arose* in (13) is tagged as 'GVM' for 'grammar-verb-morphology'.

## 2.6 Evaluation of error annotation

The usefulness of error-annotated corpora depends on the consistency of the annotation. Error annotation in learner corpora is mostly done manually and in this section we are concerned with the evaluation of manual annotation (see Chapters 25 and 26, this volume for more on automatic annotation and evaluation). Evaluation of annotation reliability is always necessary, but because there are so many possibly controversial decisions to make in error annotation (target hypothesis, tagset, error assignment, etc.), it is especially difficult and crucial to evaluate the annotation. Evaluation of annotation is done in one of two ways: either one has a gold standard (a corpus with annotations deemed to be correct) and evaluates it against this corpus,[10] or one uses several annotators to

[10] The standard measures here are recall, precision, and the f-measure; these are found in all statistics introductions, see e.g., Baayen (2008) or Gries (2009).

annotate the same subcorpus using the same tagset and guidelines and evaluates how often and where they agree (called inter-annotator agreement, inter-rater reliability, or inter-coder reliability, see Carletta 1996; Artstein and Poesio 2008). Evaluation is a necessary step in assuring the consistency of annotation. It shows which categories are clearly defined and can be assigned unambiguously, and which categories or guidelines are unclear and therefore assigned inconsistently. Evaluation is typically an iterative process – guidelines are reformulated after an evaluation, evaluated again, etc. until the result is consistent.

## 2.7 The main uses of error annotation

There are many studies that use error-annotated corpora for interlanguage research.[11] In Section 3 we highlight three studies using error annotation. At this point, we want to give an overview of the general types of error studies. Qualitative studies that focus on a small number of errors by single learners or a few selected learners are often the first step in the error analysis and give rise to hypotheses that can then be tested quantitatively. The study by Brand and Götz (2011) is a good example of this. Brand and Götz, using the error-tagged German component of the *Louvain International Database of Spoken English Interlanguage* (*LINDSEI-GE*), which is made up of spoken learner English produced by speakers with L1 German, investigate different properties of fluency and accuracy. They provide overall statistics for all learners but, crucially, they also include a detailed study of five selected learners, which leads to a deeper understanding of the interaction.

Early studies in EA sometimes counted raw error frequencies in order to find out which linguistic phenomena seemed to be especially difficult for learners. Diehl et al. (1991), for instance, motivate their corpus-based contribution about the acquisition of the German declension system with the observation that inflectional errors within the noun phrase are by far the most frequent errors among different learner groups.

Error annotations are just like other linguistic annotations, and studies using error categories can follow standard corpus-linguistic methodology. This can go far beyond the simple exploratory study by Diehl et al. (1991) into statistical hypothesis testing, multivariate analysis, and modelling (for an overview, see Gries 2008a). We cannot explain here the different methods in detail but want to structure our overview according to a simple distinction made of quantitative corpus studies introduced in

[11] There are many learner corpora with (partial or complete) error annotation – here we list only a few examples: for L2 English the *International Corpus of Learner English* (*ICLE*, Granger 2003a), the *Hong Kong University of Science and Technology Corpus* (*HKUST*, Milton and Chowdhury 1994), the *Cambridge Learner Corpus* (*CLC*, Nicholls 2003), for L2 French the *French Interlanguage Database* (*FRIDA*, Granger 2003b), for L2 German *Falko* (Lüdeling et al. 2008), for L2 Czech the *Acquisition Corpora of Czech* (*AKCES*, Hana et al. 2012), for L2 Norwegian the *Andrespråkskorpus* (*ASK*, Tenfjord, Meurer and Hofland 2006), for L2 Arabic the *Pilot Arabic Learner Corpus* (Abuhakema et al. 2008).

Biber and Jones (2009) – 'type-A studies' (pp. 1291ff.) and 'type-B studies' (pp. 1298ff.).

Type-A studies focus on one linguistic phenomenon in one corpus with the aim of understanding how different variants are distributed. Biber and Jones give the example of subordinate clauses in English, which can be introduced by *that* or nothing (*he thinks that she should smile more often* vs *he thinks she should smile more often*). The goal of a type-A study is to identify linguistic and extra-linguistic features that influence the choice between the variants. In their study they find that a combination of register, the frequency of the embedding verb, and the nature of the subject in the subordinate clause play a role. In that sense, type-A studies are detailed analyses of linguistic behaviour. Because the choice of variant is typically influenced by many factors, some of which interact, they are often described using multifactorial models. Type-A studies offer an important gain in error-based language research. Rather than looking only at errors, type-A studies can choose a linguistic phenomenon (e.g. articles in noun phrases) and count all errors (missing articles, wrong articles) as well as all correct instances of article placement in a corpus. The relation between certain erroneous structures and correct structures of the same type allows more meaningful and precise conclusions than the observation of errors without their comparison with correct cases.

Type-B studies, on the other hand, compare (counts of) categories across different corpora. Most of the recent learner corpus studies compare two (or more) learner corpora or a learner corpus and a native speaker corpus. In learner corpus research this is called Contrastive Interlanguage Analysis (CIA), see Granger (1996) and Chapter 3 (this volume).

Learner language is multifaceted, which leads to multi-method study designs (see Tono 2004; Gries 2008a; Brand and Götz 2011, and many others). For many research questions, type-A studies and type-B studies are combined. Often the analysis of error categories is combined with the analysis of other categories in the corpus.

Type-A studies are typically driven by function, not form – the basic idea being that there are several ways to express the same function. In that sense, they are variationist in nature although their purpose might be different from other (sociolinguistic or diachronic) variationist studies (cf. Labov 1978, 2004). Type-B studies compare different corpora – either different L2 corpora or a learner corpus and a native speaker corpus. Granger (2002) suggests using multiple comparisons, involving different L2s as well as L2 vs L1, in order to tease apart the different influences on learner language. Many type-B studies do not involve error tags but compare either lexical forms or annotation categories; the studies reported on here do, of course, use error annotation. Type-B studies can be cross-sectional or longitudinal. Ideally, the corpora that are being compared differ only in one extra-linguistic variable, such as L1 or proficiency level, while all other external variables are kept stable. As a result,

quantitative differences between the measured categories in the different corpora can be interpreted as an effect of the extra-linguistic variable itself. The *ICLE* subcorpora, for example, are collected according to the same criteria except for the learner's L1 and can be compared to identify L1 influences.[12]

In Section 2.3 we briefly sketched acquisition models that hypothesise that certain types of errors are typical, and possibly necessary, for a given acquisition stage. These hypotheses would have to be verified in longitudinal studies. Genuine longitudinal studies compare the same learners across different acquisition stages (ideally using the same or at least comparable tasks). Such genuine longitudinal corpora are rare and therefore sometimes replaced by quasi-longitudinal corpora in which different learner groups with different proficiency levels are investigated (see Chapter 17, this volume). A recent quasi-longitudinal study based on *ICLE* is Thewissen (2013). Thewissen aims at measuring the development of language acquisition by comparing error tags, annotated according to the *ICLE* guidelines, across different proficiency levels. In order to assess proficiency levels, the learner texts (223 essays) were rated by professional raters according to precise rating guidelines.

We want to briefly mention one problem that pertains to all type-B studies and is often ignored. Learner corpora are typically collections of texts by different learners. The corpus design specifies a number of external variables such as L1, level of proficiency, text type or mode of acquisition. The texts within the collection are then treated as a homogeneous corpus. Put in statistical terms: all texts are seen as samples from the same population. This implies that the internal grammars of all the people who contribute to the corpus follow the same system. The corpus is then compared to another corpus which differs in (at least) one design parameter. This can be statistically problematic whenever the within-group variation is too high or when there are clusters within the corpus. This shows that the samples cannot stem from one population. At the very least, variances should be reported, but often it might be necessary to calculate a model that takes such effects into account (see Evert 2006; Gries 2009 for more on this issue).

There are studies which go a step beyond classical type-A or type-B studies and combine both types in order to find out which categories are overused or underused by learners compared to native speakers. Good examples are the two papers reported in Section 3: Maden-Weinberger (2009) (Section 3.1) and Díez-Bedmar and Papp (2008) (Section 3.2). Those studies use variationist designs which are able to analyse the overuse and underuse not only of form but also of function. For studies like these one needs a corpus with (often quite specific) error tags as well as grammatical

[12] This is an idealisation. The teaching method, previously studied languages and possibly situational parameters in the collection may also differ.

information like part-of-speech information or even syntactic information (compare Hirschmann et al. 2013 for a study on the use of modification by learners of L2 German which uses a parsed corpus).

Carefully done, type-A studies or combined studies are one way to avoid the comparative fallacy. Studies that start with functions instead of forms are difficult to carry out because they have to find all places in a corpus where the function under consideration is used, which is a matter of interpreting the data. How hard this problem is depends on the phenomenon: it is not problematic to find all nouns in a corpus and see whether they are preceded by an article. It is, however, very difficult to see which sentences 'require' modality. Again, standoff architectures would be helpful and it is good practice to make the complete information available.

Type-A and type-B studies use error annotation to investigate acquisition processes. For this purpose the learner corpora are annotated independently of the learner and the learner does not ever see the annotation. Acquisition studies feed back into teaching only indirectly. But error annotation can be and has been used directly in teaching. In many settings, a learner is allowed to produce drafts of an assignment which are commented on by a teacher, corrected by the student, resubmitted, etc. until the final version is turned in for grading (see Burstein et al. 2004). A number of very interesting corpora have developed from this setting. These are treated in more detail in Chapters 20 and 22 (this volume) but we will sketch one such corpus (Section 3.3) in order to illustrate the possibilities after we have introduced two representative type-A studies.

## 3 Representative studies

**3.1 Maden-Weinberger, U. 2009.** *Modality in Learner German: A Corpus-Based Study Investigating Modal Expressions in Argumentative Texts by British Learners of German*. Unpublished Ph.D. thesis, Lancaster University.

A typical variationist study is reported in Maden-Weinberger (2009), which analyses the expression of modality by learners of German as a foreign language with L1 English. Maden-Weinberger aims to analyse to what extent modality is expressed differently by these learners in comparison to German native speakers, and which words, structures or morphemes are difficult to acquire. To achieve this objective, she collects, annotates and compares the *Corpus of Learner German* (*CLEG*), consisting of argumentative essays written by English learners of German as a foreign language, a comparable German native speaker corpus (*KEDS*, *Korpus von Erörterungen deutscher Schüler*), collected from German secondary school students, plus a native English corpus (*LIMAS*, *Linguistik und Maschinelle Sprachbearbeitung*) and a German–English translation corpus (*INTERSECT*, *International Sample of English Contrastive Texts*). The last two corpora are used to compare the L2–L1 differences with general differences in expressing modality in

native English and German. The investigation focuses on a variable or function – the expression of epistemic and deontic modality – and starts with defining possible forms to express this function, such as modal verbs, adverbs and adverbials or subjunctive mood. Maden-Weinberger annotates all forms functioning as epistemic or deontic modal expressions that can be found in her corpus data. In addition, she tags all errors regarding modality. In this way, she can see which forms are used by learners and which of these seem especially difficult. She can prove that the learners in her study use different modal expressions than the German native speakers in the comparable native speaker corpus: while the learners avoid (underuse) modal verbs, they overuse different adverbial words and phrases to express epistemic modality, although German modal verbs like *werden* 'will' are generally overused by the learners. She concludes that the epistemic function that modal verbs in German can have is especially hard to acquire by English learners.

**3.2 Díez-Bedmar, M. B. and Papp, S. 2008.** 'The use of the English article system by Chinese and Spanish learners', in Gilquin, G., Papp, S. and Díez-Bedmar, M. B. (eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi, pp. 147–75.

Díez-Bedmar and Papp (2008) analyse the use of articles in English by two learner groups whose L1s differ with respect to articles and the marking of definiteness, Chinese and Spanish. While Chinese does not use articles at all, the Spanish article system is similar to the English one, with rather subtle differences regarding pragmatic aspects. The authors analyse the misuse of articles by Chinese and Spanish learners of English in a language transfer perspective, with the hypothesis that both learner groups will produce pragmatic article errors, while Chinese learners will also produce strictly grammatical article errors. In an extensive corpus study, student essays of Chinese and Spanish learners of English as a Foreign Language are collected and analysed. Again they combine an error analysis with a CIA. Different semantic features that influence the type of article, such as genericity, definiteness and specificity of the respective noun phrase, are annotated and taken into account in the statistical analysis. The CIA shows that the Chinese learners grammatically avoid (underuse) articles in contrast to Spanish learners of English. The error analysis shows that the Chinese learners produce more errors in all semantic contexts (except for erroneous zero articles in generic uses) and that the specific pragmatic contexts are more difficult for both learner groups (indefinite articles in generic contexts show the least accurate uses for both learner groups).

Both exemplary studies above show that it can be very helpful to study one phenomenon in detail and annotate additional information (such as the type of modality or the factors influencing article choice), rather than working with a general tagset that aims at addressing many different

types of errors at once (see Meunier 1998 for a discussion of the granularity of tagsets). The studies also show that error analysis can be fruitfully combined with other methods of analysis such as CIA.

**3.3 Lee, J., Yeung, C. Y., Zeldes, A., Reznicek, M., Lüdeling, A. and Webster, J. 2015.** 'CityU corpus of essay drafts of English language learners: A corpus of textual revision in second language writing', *Language Resources and Evaluation*. doi: 10.1007/s10579-015-9301-z

Our third case study concerns the use of error annotation in teaching. In a corpus of academic L2 English collected at the City University of Hong Kong, students were allowed to submit as many drafts as they wanted before turning in the final assignment. The teachers commented on each submission and mostly used error tags from a predefined error tagset. Each student then made the corrections he or she wanted to make and resubmitted the text. Submissions and teacher feedback were collected and stored as a parallel corpus so that each version of the text can be compared with the other versions. Example (16) is a corrected version of the sentence in (15). Note that the student corrected the article (*the program → a program*) after getting an error code 'wrong article'. The verb inflection (*have improve → have been improved*) was changed but is still not correct.

(15) I learned the function of the Visual Basic and due to the debugging of the program, I *have improve* my understanding in the structure of the program code. it would be useful at next time I write **the** program.

(16) I learned the function of the Visual Basic and due to the debugging of the program, I **have been improved** my understanding at the structure of the program code. It will be useful next time I write **a** program.

Corpora like these can be used to assess what effect an error code has on the student and how many errors are actually corrected and how. Studies that use error-annotated corpora in this way include Wible et al. (2001) and O'Donnell (2012a).

## 4 Critical assessment and future directions

Error-annotated corpora make it possible to investigate whether, for example, learners of L2 German underuse modification or whether they just underuse specific types of modifiers, how speech rate and accuracy interact in L2 English, or which errors appear and disappear in which stage of acquisition. We believe that there are still many open methodological and conceptual issues in the study of learner corpora. Some of these (corpus design, acquisition processes, statistical modelling) are discussed in other chapters of this handbook. The open issues that pertain to error

annotation include reproducibility and replicability, the interpretation of errors, and the combination of error studies and other learner data.

If learner corpora are available with error annotation and if error annotation is done in a transparent and clear way, error studies become reproducible and results become replicable. In this chapter we showed that error annotation always depends on an interpretation of the data. Bley-Vroman (1983) warns against analysing one variety (the learner language) through the eyes of another variety (the 'standard' of the target language) and says that the properties of learner language can only be understood if the L2 variety is studied as a genuine variety in and of itself. While Bley-Vroman is certainly right, it is also clear that many of the properties of learner language can only be understood if learner language is compared to target language structures. The first step in error annotation is the reconstruction of a target grammar utterance – called target hypothesis – against which the learner utterance is evaluated. There can be many such target hypotheses for a given learner utterance. Whichever one is chosen in a given corpus influences the error exponent and the error tag – and the analysis that follows from these. We argued that it is useful to make the target hypothesis explicit and to use a corpus architecture that allows multiple target hypotheses. Error annotation studies often combine the evaluation and counts of the error tags with other corpus information such as part-of-speech tags, syntactic analysis or statistical patterns of lexical information within the corpus. Most of these studies combine several methods, which helps minimise the comparative fallacy.

Another interesting issue is that of standardisation of error tagsets. While some degree of standardisation in terms of edit-distance-based tagging (as described above) is useful, there is some doubt as to the desirability of more fine-grained standardisation. The scope and granularity of an error tagset depends on the phenomenon to be studied and the research question to be answered. In flexible corpus architectures it is possible to add one or several layer(s) for the fine-grained analysis of a given error type or phenomenon. The development of error tagsets for a given phenomenon could be viewed as the most important step in understanding it (and thus is an integral and necessary part of research).

Because it involves so many decisions, manual error annotation is time consuming. In the future we will see more and more semi-automatic and automatic methods for error annotation (see Chapter 25, this volume). It is especially important to test and report the reliability of error annotation, be it manual or automatic. There are different ways of testing reliability. Manual annotation is typically tested by comparing the decisions made by two or more annotators (inter-annotator agreement, inter-rater reliability, see Section 2.6), while automatically annotated corpora are typically evaluated by comparing the results with a training corpus.

In the future, we will also see even more statistical modelling of errors and other properties in learner corpora (see Chapter 8, this volume). This corresponds to the trend in grammar and acquisition models – away from categorial, algebraic models to probabilistic, usage-based models.

## Key readings

Corder, S. P. 1967. 'The significance of learner's errors', *International Review of Applied Linguistics in Language Teaching* 5(1–4): 161–70.

This book is the first and still very useful approach to integrating the notion of learner errors into a comprehensive analysis of learner language and theory of second language acquisition. It contains the basic concepts of linguistic errors that were discussed in this chapter, argues for the necessity of errors in the language acquisition process, and discusses similarities and differences between the first language and second language acquisition process.

Lennon, P. 1991. 'Error: Some problems of definition, identification, and distinction', *Applied Linguistics* 12(2): 180–96.

This article defines and discusses basic concepts in error analysis and fundamental distinctions of error types. This is exemplified by a learner corpus study of advanced English learners.

Ellis, R. and Barkhuizen, G. 2005. *Analysing Learner Language*. Oxford University Press.

This book introduces and discusses the essential methods for a comprehensive analysis of spoken and written learner language. Chapter 3 is dedicated to error analysis, providing a historical and theoretical background and leading the reader through the different steps of a state-of-the-art error analysis.

Díaz-Negrillo, A. and Fernández-Domínguez, J. 2006. 'Error tagging systems for learner corpora', *Revista Española de Lingüística Aplicada* 19: 83–102.

The article provides an overview of existing error taxonomies. Díaz-Negrillo and Fernández-Domínguez compare different learner corpora implementing error classifications and discuss the conceptual differences of the approaches.

Granger, S. 2008b. 'Learner corpora', in Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: Mouton de Gruyter, pp. 259–75.

In this article, Sylviane Granger explains the basic methodology of using learner corpora in the study of second language acquisition. Alongside other essential methods in learner corpus research, she describes different aspects of error annotation and how it can be used in acquisition studies, in computer-assisted language learning and in teaching.

Dagneaux, E., Denness, S. and Granger, S. 1998. 'Computer-aided error analysis', *System* 26(2): 163–74.

The paper explains how EA problems can be overcome by using error-annotated corpora, introducing data from the *International Corpus of Learner English* and the error tagset used in the corpus.

Reznicek, M., Lüdeling, A. and Hirschmann, H. 2013. 'Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture', in Díaz-Negrillo, A., Ballier, N. and Thompson, P. (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: Benjamins, pp. 101–23.

The authors argue for explicit and multiple target hypotheses in a multi-layer corpus architecture. In addition to the methodological problem of deciding on one target hypothesis, they show that different target hypotheses (and, based on these, different error tags) highlight different types of errors.