Zeitschrift für Interkulturellen Fremdsprachenunterricht
Didaktik und Methodik im Bereich Deutsch als Fremdsprache

ISSN 1205-6545 Jahrgang 19, Nummer 2 (Oktober 2014)

Towards an Error Taxonomy for Student Writing

Nikola Dobrić

Department of English and American Studies
Alpen-Adria-Universität Klagenfurt, Austria
E-mail: Nikola.Dobric@aau.at

Guenther Sigott

Department of English and American Studies
Alpen-Adria-Universität Klagenfurt, Austria
E-mail: Guenther.Sigott@aau.at

Abstract: Current practice in writing assessment has tended to divert attention from the notion of error in favour of more global and intuitive descriptions of learner performance. As a result, experts tend to disagree in the way errors in student writing are described. This brings about complications for student feedback as well as for studying the construct validity (cf. Messick 1989) of rater-mediated assessments of student writing by means of computer-corpus-based methodology, which requires reliable annotation. In order to alleviate these complications, an error taxonomy is proposed which could serve as a basis for student feedback on the one hand, and as a basis for corpus-based studies into the construct validity of large-scale assessments of writing, on the other.

Die gegenwärtige Praxis der Beurteilung von Kompetenz im Schreiben tendiert dazu, den Begriff des Fehlers zugunsten von eher globalen und intuitiven Beschreibungen von Lernerperformanz in den Hintergrund zu drängen. Dies führt unter ExpertInnen zu Uneinigkeit bei der Beschreibung von Fehlern in Lernerperformanzen. Einerseits entstehen dadurch Komplikationen bei der Formulierung von Lerner-Feedback. Anderseits entstehen Komplikationen bei Untersuchungen zur Konstruktvalidität (cf. Messick 1989) von beurteilergestützten Testverfahren mittels computerbasierter Methodologie, die zuverlässige Fehlerannotierung voraussetzt. Zur Abhilfe wird hier eine Fehlertaxonomie vorgeschlagen, die sowohl die Formulierung von Lerner-Feedback erleichtern als auch bei computerbasierten Validierungsstudien von Kompetenzbeurteilungen im Schreiben mit großen ProbandInnenzahlen hilfreich sein soll.

Key words: Error analysis, error taxonomy, learner corpora, scope, substance, annotation, inter-rater agreement, inter-annotator agreement, rating, validation, feedback

1. Introduction

Before the advent of the Common European Framework of Reference for Languages (CEFR), assessments of written production in L1, L2 or FL settings in Europe tended to be based either on error counts or on necessarily subjective intuitive judgements made by the individual teacher or examiner (see, e.g., Spolsky 1995: 59-63, 322-326 for English FL and L2 writing; Sigott (personal experience in the Austrian school context for L1 German as well as L2 and FL English and French writing)). At times, both approaches were combined to yield assessments in which it remained unclear how much each of the two approaches contributed to the final grade. The CEFR attempts to offer students, teachers and examiners a common language to describe levels of attainment at each of six levels of proficiency. The descriptors for the individual levels are informed by a philosophy that focuses predominantly on positive features while negative features are usually only referred to indirectly, if at all, as the following example shows, where the relevant formulation is in bold typeface:

> Can summarise, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his/her field **with some confidence** (CEFR scale for Reports and Essays, Level B1).

Negative features are explicitly referred to only in the scales for qualitative aspects of spoken language use (Council of Europe 2001: 28-29), vocabulary control (ibid. 112) and grammatical accuracy (ibid. 114):

> Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes (CEFR scale for Qualitative Aspects of Spoken Language Use, Level B2).
>
> Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations (CEFR scale for Vocabulary Control, Level B1).
>
> Uses some simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say (CEFR scale for Grammatical Accuracy, Level A2).

In classroom practice, the absence of negative features in most of the can-do statements sometimes turns out to be problematical when it comes to giving students feedback.

Inherent in the CEFR is a preference for judging rather than counting in the practice of assessment. Accordingly, both positive and, to a limited extent, negative features are described in holistic terms in order to make the descriptors applicable to as wide a range of performances and useful to as large a range of users as possible. This approach has helped in making the assessment of language proficiency more uniform, reliable and possibly valid. However, it has diverted attention from the notion of error as a manifestation of learner language at a particular state of development towards full proficiency. As a result, teachers and researchers will more often than not describe the same error differently. This has consequences for our ability to provide construct validity evidence (cf. Messick 1989; see Kecker 2011 for an overview) for assessments of written production as well as for the kind of feedback (see Reitbauer, Mercer, Schumm-Fauster & Vaupetitsch 2013 for an overview) that learners receive.

In certain assessment situations it is important to go beyond stating that certain aspects of learner performance do not conform to the norm. In addition to this, learners expect to be told why their performance is not norm-adequate, and what they would need to consider if they wanted to develop their proficiency further to approximate the norm better. Learners need feedback that enables them to employ strategies for development. Learners need information about what it is that makes a particular instance of performance wrong. Clearly, such feedback is only valuable if it is given in terms that can be easily understood by learners and that will not differ depending on the person who gives the feedback. Experience shows that ways in which feedback is provided with regard to the same performance differ widely among teachers and experts. Interestingly, little attention has been paid to ways in which errors can be described in more standardised ways that will not vary from expert to expert.

The extent to which assessments of the productive skills in CEFR terms depend on the number of errors is an open question which needs to be addressed in any attempt to validate scores resulting from the application of writing scales. Such an approach presupposes a stable and replicable error count which does not depend on the individual expert identifying the errors. If learner performances are to be kept in a computer corpus to provide a data base for analysis, the annotation of the corpus in terms of errors is only meaningful if it is based on an error taxonomy that will yield similar results across different annotators. The error taxonomy that is described below is intended to be useful both as a basis for validating assessments of written skills and as a basis for feedback to learners in teaching practice.

2. Problems with existing error classifications

Error taxonomies, when used as a basis for effective feedback, should be plausible to learners and teachers without giving rise to issues of interpretation. Similarly, when used as a basis for corpus annotation for the purpose of large-scale validation studies, they are supposed to yield reliable results. However, a standardized error taxonomy for language teaching and testing or for processing learner-language corpora does not exist (cf. Díaz-Negrillo & Fernández-Domínguez 2006: 98). Instead, there is a multitude of error taxonomies, and this diversity becomes particularly noticeable when it comes to processing learner corpora. As a result, feedback on the same student performance will usually differ from teacher to teacher, and corpus annotators will disagree on the way they tag the same performance unless they have undergone extensive training in using a particular error taxonomy.

In learner corpus annotation, an error taxonomy will be the more useful the less its application and the error tagging resulting from it is influenced by idiosyncrasies of individual annotators. Given the large amount of data that needs to be processed in learner corpus development, it is usually necessary to recruit help to process the data manually (or semi-automatically). Clearly, the more precisely the errors are defined and classified, the faster and more accurate the annotators will be. The level of inter-annotator agreement will depend on the ease with which errors can be identified and classified. The less the application of an error taxonomy is influenced by annotator variables, the more useful it will be for corpus annotation (cf. Carletta 1996: 250-252; Gwet 2001; Landis & Koch 1977: 160; Viera & Garrett 2005: 360).

Error taxonomies, when used as the basis for corpus annotation, then, are supposed to yield reliable results. At the same time, they should also be plausible to learners and teachers so that they could be used as a basis for effective feedback. Different approaches to error classification have been used in attempts to satisfy these requirements:

(1) classification according to level of linguistic description: this most commonly applied error taxonomy employs the various levels of linguistic analysis (phonology, morphology, syntax, semantics, etc.) as the basis for defining error types (see, e.g., Ellis & Barkhuizen 2005; George 1972; Havranek 2002). Such taxonomies identify errors such as ‘passive voice’, ‘temporal conjunctions’, ‘transitive verbs’ or ‘wrong word’;

(2) classification according to alterations in ideal performance: this less popular and more abstract type of taxonomy describes errors in terms of what has been altered on the ‘surface’ level of a hypothetical ideal performance so that the learner performance came about. This includes *omissions* (some element demanded by the norm is left out), *additions* (some element barred by the norm is added), *misinformation* (some element is expressed by a form barred by the norm), and *misordering* (elements are ordered in a manner barred by the norm) (see, e.g., Dulay, Burt & Krashen 1982: 150);

(3) classification combining level of linguistic description and alterations in ideal performance: this approach describes each error both with regard to level of linguistic analysis and in terms of alteration in the hypothetical ideal performance, thus yielding error categories like *tense/omission* or *modal verbs/misordering* (see Pibal 2012);

(4) classification in terms of presumptive cause of error: this approach attempts to describe errors with regard to the presumptive source of the error. This may be the learner’s L1 or another foreign language, or universal cognitive constraints. Error categories in such taxonomies are interlingual errors (attributable to interference), developmental errors (due to universal cognitive constraints), ambiguous errors (attributable to more than one possible source), and unique errors (a residue category for unclassifiable errors) (cf. Dulay et al. 1982: 163); and

(5) classification according to the degree of message impairment: this approach describes errors in terms of the degree to which they disturb the message in information theory terms. Errors are here characterized with regard to their effect on the listeners or readers. A distinction is often made between global errors and local errors. Global errors involve large amounts of noise and seriously impair comprehensibility. An example would be violations of major syntactic rules. Local errors are said to cause noise to a lesser degree and involve a narrower focus. Examples are errors in article use or verb inflections (cf. Dulay et al. 1982: 172).

However, attempting to apply these taxonomies to authentic learner language data has shown us that they sometimes leave room for a considerable amount of subjective judgment, which in turn renders the training of future corpus annotators difficult. Also, reports on such training are hard to find in the literature, and, to the best of our knowledge, there have been no systematic studies of the effects of such training. This has prompted us to devise an approach to error classification which we hope will leave less room for subjective judgment, therefore prove less demanding in annotator training, and produce high inter-annotator agreement. We also believe that apart from corpus annotation, our taxonomy should have the potential to serve as a basis for more uniform feedback to students, which will be the subject of a future publication.

3. Description of the Model

The model is based on the grammatical hierarchy described in *A Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech & Svartvik 1985). Central to the model is the distinction between *scope* and *substance*. *Scope* refers to the amount of textual or extratextual context that is required for recognising the presence of an error. *Substance*, by contrast, refers to the size of the element that needs to be changed in order to correct the error.

3.1. Scope

The *scope* of an error may be *word, phrase, clause, sentence* or *text*. This is demonstrated in examples 1 to 7. The examples are gleaned from writing performances produced by fourteen-year-old Austrian pupils in the context of the Austrian National Educational Standards for English and from writing performances by students of English in writing classes of the Department of English and American Studies at Klagenfurt University.

[Example 1] We learn a lot about the *atraktions*.

Here it is obvious that ‘atraktions’ deviates from the norm of English spelling. The lexical item that is targeted is ‘attractions’. A fully proficient speaker of English would most probably recognise the error even if the word stood in isolation without any context at all. We will therefore say that the *scope* of this error is *word*.

[Example 2] *At evening* I and my best friends always speak *at our problems and other things*.

Unlike in example 1, here the errors become apparent only when one looks beyond the individual words. While ‘at’ and ‘evening’ are both perfectly acceptable as lexical items of English, the combination ‘At evening’ violates the norm. So, unlike in example 1, the error only becomes manifest when one takes into consideration the context beyond the individual word. In cases like these, we will say that the *scope* of the error is *phrase*.

[Example 3] *There was it beautiful* and very interesting.

In example 3 the error only becomes noticeable when one extends the scope beyond the verb phrase to the level of the clause. Only when ‘beautiful’ is added to the scope does it become clear that clause structure rules of English have been violated. Example 2 contains a similar case. While ‘at our problems and other things’ is a perfectly grammatical construction, adding ‘speak’, also a perfectly acceptable English word by itself, makes the construction unacceptable. In cases like this, we will say that the *scope* of the error is *clause*.

[Example 4] There is a lot of evidence that body art was used three to five thousand years BC, and it is believed that the first *one of them* was made by accident.

Example 4 shows a sentence consisting of two coordinated clauses. When considered in isolation, neither of them violates any grammatical rules of English. However, when they are combined, it becomes obvious that ‘body art’, being an uncountable noun, cannot serve as an antecedent for ‘one of them’, which presupposes a countable antecedent. So in this case it is not enough to consider clause-level context, but it is necessary to widen the scope to the level of the sentence. In cases like this, we will say that the *scope* of the error is *sentence*.

[Example 5] Yesterday our class was in the "House of Music" in Vienna. There you can see lot of things of Mozart or Beethoven. Most of the time we were listening to strange noises. It was very great fun. Some stations were a bit boring, but it was okay. It was very interesting and next week we are going to have a test about classical music. I hope I *did* that well. The funniest thing was that we missed the train back home. So we had to wait two hours. In the meantime we went shopping. I didn't know that there are such great shopping centers in Vienna. I liked it very much there because it's very interesting and exciting. You can learn a lot there.

Example 5 shows how an error only becomes noticeable when the scope is widened beyond sentence boundaries. The simple sentence 'I hope I did that well.' only becomes problematical when the scope is widened beyond sentence boundaries. Then it becomes clear that 'that' refers to 'test' in the preceding sentence, and in the time frame of the text this test is located in the future. In cases like this, we will say that the *scope* of the error is *text*.

3.2. Substance

The *substance* of an error refers to the smallest constituent in the learner production that needs to be modified so that the error will disappear. Like *scope,* it is described by recourse to the grammatical hierarchy in Quirk et al. (1985). Like *scope,* the *substance* of an error may therefore be *word, phrase, clause, sentence* or *text.*

In examples 1 (repeated here), 6, 7, 8 and 9 a change in a single word is sufficient to remove the error.

[Example 1] We learn a lot about the *atraktions*.

In order to rectify the error in example 1, the orthographic shape of the word needs to be changed from 'atraktions' to 'attractions'. We will therefore say that the *substance* of the error is *word*. In this case, *scope* and *substance* are at the same level, namely *word*.

[Example 6] We have *get* there.

The error in example 6 becomes noticeable when we look at the verb phrase 'have get'. Correcting it involves a change at word level. Either 'have' is replaced by a different modal verb, e.g., 'could', or 'get' is changed to 'got'. In either case, the change is at word level, so we will say that the *substance* of the error is *word* while the *scope* is *phrase*.

[Example 7] *Were* is my homework book?

In example 7 the error becomes apparent when we take into account clause-level context. So the *scope* of the error is *clause*. However, the *substance* of the error is *word* because correcting it involves changing 'Were' to 'Where'.

[Example 8] One solution to the problem of too high alcohol consumption is to simply not authorize everyone to sell *it*.

Example 8 contains an error concerning 'it'. Looking at the elliptical clause 'to simply not authorize everyone to sell it.' is not enough for the error to become visible. Only when the *scope* is extended to the entire sentence does it become clear that there is no suitable antecedent for 'it'. 'it' needs to be replaced with a different word, e.g., 'alcohol' for the error to disappear. Thus, the *scope* of the error is sentence while the *substance* is word.

[Example 9] I *arrive* there with a long delay.

In example 9 the error only comes to the fore when one looks beyond the clause and sentence boundary in the learner text. The wider context from which this sentence is taken shows a narrative orientation of the text and it becomes clear that 'arrive' should read 'arrived'. Thus, a change at word level will make the error disappear. We will therefore say that while the *scope* of the error is *text*, the *substance* is *word*.

Examples 10 to 13 require a change of phrase structure in order to correct the error.

[Example 10] Our teacher expects us to know a lot about the historical background *of the covered literary epoch*.

In example 10, a look at the noun phrase ‘the covered literary epoch’ is enough for us to become aware of an error, which consists in ‘covered’ being used in premodifying rather than in postmodifying position. The phrase structure in the learner performance will need to be changed to a structure in which the head is followed by a post–modification rather than preceded by a premodification. So a change at phrase level is required in order to remove the error. Here both *scope* and *substance* of the error will therefore be said to be *phrase*.

[Example 11] When you *will* phone me, I’ll come.

In example 11 the error becomes apparent when the *scope* is widened to clause level. Then it becomes clear that ‘will’ is not acceptable in a temporal clause introduced by ‘when’. ‘Will’ needs to be removed for the error to disappear. This involves a change in phrase structure from a verb phrase containing an auxiliary slot to one without such a slot. Consequently, we will say that while the *scope* of the error is *clause*, the *substance* is *phrase.*

[Example 12] The thieves observed him when they *have stolen* his wallet.

Example 12 demonstrates an error for which the *scope* is *sentence* while the *substance* is *phrase.* While the two clauses are each acceptable when considered in isolation, their combination renders ‘have stolen’ unacceptable. The verb phrase ‘have stolen’ would need to be changed to *stole* if the error was to be corrected.

[Example 13] *The* parliament eventually passed this contentious bill. [from a text about the British Parliament]

In example 13 it only becomes apparent that the determiner ‘The’ is problematical when context beyond the sentence boundaries is considered. The structure of the noun phrase ‘The parliament’ needs to be changed by removing the determiner slot. We will therefore say that while the *scope* of this error is *text*, the *substance* is *phrase.*

In example 14 below it is obvious that the order of ‘put he’ is at fault. The error becomes noticeable at clause level, hence the *scope* is *clause.* Since correcting it will involve changing the structure of the clause, the *substance* is also *clause.*

[Example 14] Then *put he* down his skis and went into the hut.

[Example 15] Only when we save money ahead of time *we will* be able to afford a family holiday.

Example 15 demonstrates a classic error in English. While ‘we will be able to afford a family holiday’ is a grammatically acceptable English clause, it becomes unacceptable when the scope is widened to the entire sentence. Then it becomes clear that the order of the constituents in the verb phrase ‘we will’ needs to be reversed to ‘will we’. So we will say that while the *scope* is *sentence,* the *substance* is *clause.*

Examples for the combination *scope: text, substance: clause* or *sentence* are less frequent than the ones discussed so far. They mostly involve sequencing problems to do with theme – rheme structure in discourse. Example 16 is a case in point.

[Example 16] It is London that is the capital of England. [beginning of a text about London]

Here it takes context beyond the sentence for the error to become apparent. Removing the error would involve changing the sentence structure to *London is the capital of England.* Consequently, the *scope* will be *text* while the *substance* will be *sentence.*

Examples for the combination *scope: text, substance: text* typically violate conventions of text structure like topic sentence – supporting detail – concluding sentence. In these cases, while the problem only becomes apparent when one looks beyond sentence boundaries, rectifying the problem involves changing the order of the sentences in the text.

4. Outlook: Applications in teaching and in validation research

We believe that the model described in this article is useful for teaching practice as well as for studying the construct validity (cf. Messick 1989) of assessments of writing skills. In teaching, the model has potential for making students aware of the complex notion of textual competence. The distinction between *scope* and *substance* should help students understand that it is not enough to pay attention to lexical items or syntactic structures per se, but that it is the effect that the use of such structures has on the entire fabric of a text that needs to be taken into account as well. The model should therefore be a useful tool for making students aware of the phenomena of cohesion and coherence in texts.

In validation research, the model provides a basis for more reliable error counts. Such error counts can be correlated with rater-mediated assessments of performances and shed light on what phenomena raters pay particular attention to in the rating process. If performances are stored in a computer corpus, the model is expected to be a good basis for corpus annotation in terms of errors as it should be easy to train annotators to use the model in reliable ways. Currently it is being piloted in a project aimed at studying the contribution of norm-deviant features to rater-mediated assessments of writing skills in the Austrian National Standards Tests. In this first phase, the focus is on studying annotator agreement (Sigott, Cesnik & Dobric, forthcoming).

References

Carletta, Jean (1996), Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22: 2, 249-254.

Council of Europe (2001), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Díaz-Negrillo, Ana & Fernández-Domínguez, Jesús (2006), Error tagging systems for learner corpora. *RESLA* 19, 83-102.

Dulay, Heidi; Burt, Marina & Krashen, Stephen (1982), *Language Two*. Oxford: Oxford University Press.

Ellis, Rod & Barkhuizen, Gary (2005), *Analyzing Learner Language*. Oxford: Oxford University Press.

George, Vernon (1972), *Common Errors in Language Learning: Insights from English*. Rowley: Newbury House.

Gwet, Kilem (2001), *Handbook of Inter-rater Reliability*. Maryland: STATAXIS Publishing Company.

Havranek, Gertraud (2002), *Die Rolle der Korrektur beim Fremdsprachenlernen*. Frankfurt/Main: Peter Lang.

Kecker, Gabriele (2011), *Validierung von Sprachprüfungen. Die Zuordnung des TestDaf zum Gemeinsamen Europäischen Referenzrahmen für Sprachen*. Frankfurt/Main: Peter Lang.

Landis, Richard & Koch, Gary (1977), The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

Messick, Samuel (1989), Validity. In: Linn, Robert L. (ed.), *Educational Measurement.* Third edition. New York: Macmillan.

Pibal, Florian (2012), *Identifying errors in the written manifestations of Austrian English learner language at 8th-grade secondary level and their influence on human ratings*. MA Thesis. Alpen-Adria-Universität Klagenfurt, Austria.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey & Svartvik, Jan (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Reitbauer, Margit; Mercer, Sarah; Schumm-Fauster, Jennifer & Vaupetitsch, Renate (eds.) (2013), *Feedback Matters.* Frankfurt/Main: Peter Lang.

Sigott, Guenther; Cesnik, Hermann & Dobrić, Nikola (forthcoming), The SD Taxonomy and rater agreement. A validation study.

Spolsky, Bernard (1995), *Measured Words*. Oxford: Oxford University Press.

Viera, Anthony & Garrett, Joanne (2005), Understanding interobserver agreement: The Kappa statistic. *Family Medicine* 37: 5, 360-363.