Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture

Marc Reznicek, Anke Lüdeling, Hagen Hirschmann
Humboldt-Universität zu Berlin

Error annotation is a key feature of modern learner corpora. Error identification is always based on some kind of reconstructed learner utterance (target hypothesis). Since a single target hypothesis can only cover a certain amount of linguistic information while ignoring other aspects, the need for multiple target hypotheses becomes apparent. Using the German learner corpus Falko as an example we therefore argue for a flexible multi-layer standoff corpus architecture where competing target hypotheses can be coded simultaneously. Surface differences between the learner text and the target hypotheses can then be exploited for automatic error annotation.

Keywords: target hypothesis, multi-level corpus architecture; automatic error annotation; Falko learner corpus

1 Introduction: Why corpus architecture matters

While a lot of work in learner corpus linguistics has focused on the corpus design (for references see e.g. Granger 2008) not much attention has been paid to the corpus architecture. This is unfortunate because the underlying data model and the corpus architecture technically determine the ways in which a corpus can be used. In our paper we argue that for special, relatively small corpora that represent 'non-standard' language such as learner corpora it is very valuable to have a multi-layer standoff architecture in which all annotation layers are represented independently of each other. Standoff architectures make it possible to represent different annotation formats (tokens, spans, trees etc.) as well as enabling the user to add annotation layers at any point. They thus ensure maximal flexibility when dealing with data for which an interpretation is difficult and often controversial.

Our arguments in this paper focus on the need for target hypotheses in learner corpora. In Section 2 we show that adding an explicit target hypothesis is necessary for transparent analysis and all kinds of further annotation of learner corpora but that it is nearly impossible to agree on *one* target hypothesis for a learner utterance. It is therefore useful to provide a corpus architecture that allows the addition of several, possibly conflicting target hypotheses. We will then (Section 3) illustrate our arguments

with a detailed study of competing target hypotheses in the German learner corpus Falko.

2 What kind of information should a learner corpus provide and what kind of data is needed?

Learner corpus studies typically use one of two major methods: Contrastive interlanguage analysis (CIA) or error analysis (EA) (Granger 2008). Both methods assume that learners possess a systematic internal grammar, called interlanguage (Selinker 1972), which can be explored by looking at (naturally occurring) learner utterances and that learner corpora are one source of relevant data. CIA (see e.g. Aarts, Granger 1998; Abe 2004; Belz 2004, Tono 2004) looks at patterns in learner language by comparing categories (such as words, part-of-speech categories etc.) in learner corpora with categories in other corpora (such as native speaker corpora). It is typically quantitative. EA (see e.g. Dagneaux et al. 1998; Weinberger 2002; Izumi, Isahara 2004; Crompton 2005, Chuang, Nesi 2006), on the other hand, classifies and analyses learner errors. CIA and EA lend themselves to different research questions and operate on different kinds of data but both need interpreted (annotated) data. Generally, CIA can be done on any kind of linguistic category (lexical, morphological, syntactic, or text-based) that is annotated in the corpus. EA requires specific error annotation (see Díaz-Negrillo, Fernández-Domínguez 2006 for an overview of error tags) which can pertain to errors on any linguistic level (word, phrase, sentence etc.). The acquisition and coding of a learner corpus is typically very time consuming and expensive and it is therefore desirable for a learner corpus to be usable for many research questions.

In principle corpus annotation can be stored

a. in a tabular format where annotation is connected to tokens. Tabular formats are used for many large corpora because they allow fast indexing and search. It is possible to add further token-based annotations layers but it is not possible to add span-annotation or graphs.
b. in a tree (XML or otherwise) which allows token and span annotations as well as hierarchical annotations, but not graphs or conflicting hypotheses. Tabular formats and tree formats are inline formats, i.e. the annotations are stored in the same file as the original data.
c. in a standoff format where each annotation layer is stored separately from the original text.

Most learner corpora that we are aware of use an inline architecture. In the following we want to show that this prevents re-use for questions that the original corpus designers have not foreseen and that only standoff formats are flexible enough to make free re-use of the corpus and complete transparency of the analysis possible.

2.1 POS & lemmas

Contrastive analysis can be done on the surface forms of a learner text, but for many research questions it is necessary to have part-of-speech or base form (lemma) information for every token. Automatic taggers like the tree tagger (Schmid 1994) regularly achieve an accuracy of more than 95% for newspaper texts. Learner language is problematic for automatic taggers and there are not many studies on the accuracy of tagging learner language (an exception is van Rooy, Schäfer 2002; see also Díaz-Negrillo, Fernández-Domínguez 2006). Nevertheless many learner corpora are tagged for POS and lemmatised. Both POS tags and lemmas are token-based annotations. In principle this kind of information can be stored in a tabular fashion (inline), in tree structures (XML) or in a stand-off format.

2.2 Target hypotheses

EA can take advantage of the POS tags and lemmas but it primarily needs error annotations that to a large extent have to be added manually. Many learner corpora therefore provide some kind of error annotation.[i] Error annotation is problematic because the definition of an error itself is problematic.[ii] But no matter what error definition is used it is clear that an error can only be annotated if a 'correct' version of the utterance is assumed. Following (Ellis 2009: 50) we call this implicit 'correct form' the target hypothesis (TH). Many learner corpora provide only the error tags and leave the target hypothesis implicit. Other learner corpora such as ICLE2 (Granger et al. 2009) or FRIDA[iii] offer a partial target hypothesis for the error annotated tokens but do not discuss how the target hypothesis is constructed, implicitly assuming that there is an unambiguous way of finding it and in turn the errors that result from it.

That this is not the case has been discussed in many papers (see e.g. the discussion in Tenfjord et al. 2006). A recent empirical study (Lüdeling 2008) asked five practicing teachers of German as a Foreign Language to annotate errors in several sentences and to write out their underlying target hypotheses for the entire sentences. The comparison of their results shows that error counts and error types differ considerably from one person to the next and that those differences are due to the different target hypotheses (there was not a single sentence where all five annotators agreed on a target hypothesis).

This means that we have to assume several (competing) target hypotheses for a given learner utterance (1a). In principle there is no limit to the number of possible target hypotheses. We want to illustrate this in (1)[iv] where (1b-1g) represent different possible target hypotheses for the learner utterance. While on a purely orthographic level (1b) TH might differ from learner text (LT) for the tokens "80", "woh", "Tenniswoman" and "," a grammatical TH (1c) might want to include corrections for the miss-

ing article “a” before “tennis woman” as well. Every further level (1d-g) is still more different from the original data.

(1a) LT: One can still remember Billie Jean King, **woh** was **Tenniswoman** in the **80**, and who fought for a free homosexuality.

(1b) $TH_{ORTHOGRAPHY}$: One can still remember Billie Jean King, who was tennis woman in the 80s, and who fought for a free homosexuality.

(1c) $TH_{GRAMMAR}$: One can still remember Billie Jean King, who was a tennis woman in the 80s, and who fought for one free homosexuality.

(1d) TH_{LEXIC}: One can still remember Billie Jean King, who was a tennis player in the eighties, and who fought for a free homosexuality.

(1e) $TH_{INFORMATION\ STRUCTURE}$: One can still remember Billie Jean King, who in the eighties was a tennis player in the 80s, and who fought for a free homosexuality.

(1f) $TH_{STYLE\ 1}$: One might still remember Billie Jean King, who in the eighties was a tennis player in the 80, and fought for a free homosexuality.

(1g) $TH_{STYLE\ 2}$: One might still remember the tennis player Billie Jean King of the eighties, who was a tennis player in the 80, and fought for a free homosexuality.

Since there is no single ‘true’ target hypothesis and since EA results depend so crucially on the TH, target hypotheses have to be *explicitly* given in the corpus, so that researchers can control and understand the decisions that have been made – we illustrate this further in Section 3.1. The target hypotheses must be constructed on the basis of an annotation manual which ensures that different annotators make the same decisions over a large amount of text. This manual must be publicly available. Since the usefulness of a target hypothesis can be evaluated only against a given research question, it has to be possible to add more than one target hypothesis to the same learner utterance. Unlike POS tags, target hypotheses and error annotations cannot be stored in a simple tabular format because changes and errors do not always pertain to one token and because errors might be nested inside each other. Nevertheless, most existing learner corpora use inline architectures, i.e. they store error tags (and any other annotation) in the same file as the primary data (the learner utterance). Here we want to describe the consequences that model has (see also Lüdeling 2007).

Error exponent

Some learner corpora add error tags directly after the word or sequence that contains the error. Example (2) shows the C-LEG token-based annotation model.

(2) Zum Beispiel sie **<GrVrWoMa>** sind ein bißchen rebellisch
For instance they are a bit rebellious
For instance, they are a bit rebellious.

Gr =grammatical error, Vr=Verb, Wo=word order, Ma=main clause (Weinberger 2002:29)

(2) is problematic because there are two constituents before the finite verb which is usually not permitted in German syntax.[v] Either of the two constituents ([*zum Beispiel*]$_{PP}$, [*sie*]$_{NP}$) could be there, the other one would have to be moved after the finite verb. This means that there are at least two possible target hypotheses - Weinberger's error tag here is undecided. But independent of the decision for one or the other target hypothesis this format is unsuitable because the error exponent is not structurally marked and cannot be retrieved automatically. It is not clear whether the tag pertains to the NP or to the NP plus the PP.

Conflicting spans

Many learner corpus architectures solve the marking problem by using tags that enclose the error exponent. One such model is applied in the ICLE Corpus (Dagneaux et al. 1998). Here the error exponent (italic) is framed by the error tag on the left and a target form on the right (both in bold), cf. (3).

(3) There was a forest with dark green dense foliage and pastures where a herd of tiny **(FS)** *braun* **$brown$** cows was grazing quietly, **(XVPR)** watching at **$watching$** the toy train going past.

FS= formal spelling error, XVPR=Lexico-grammatical error for verb and preposition (Dagneaux et al. 1998:166)

ICLE uses a proprietary format but XML corpora such as FRIDA (Granger 2003) or the Corpus of Japanese Learner English NICT JLE (Izumi et al. 2004) enclose the error exponent in a similar fashion, as shown in (4) where the token "team" is annotated as a number error on a noun. Inside the XML tag the corrected form (target hypothesis) "teams" is displayed.

(4) I belong to two baseball <n_num crr="**teams**">*team*</n_num>.

n_num= number error on a noun, crr = corrected form (Izumi et al. 2004:121)

These formats clearly delimit the error exponent and provide an explicit target hypothesis. Inline annotation models using XML tags are more flexible than purely tabular formats but they have two major problems.[vi] First they cannot consistently describe crossing annotation tags and even more importantly it is not easy to model annotations which describe features of the target hypotheses themselves. Consider Table 1 where complex noun phrases have been annotated once for the original learner text (LT) and once for a target hypothesis (TH). The different word order in

the target hypotheses leads to a different extension of the NP span. Both spans partly overlap but neither is fully included in the other.

[TABLE 1]

(5) shows the example in Table 1 in XML representation. In the underlined part the second span opens before the first is closed. This is not allowed in standard XML.[vii]

(5) weil er <NPLT><ET1> **#die$** *ø* </ET1 > <NPLT2>Ziele, <ET2> **#die wichtiger als ich sind**</NPLT>, **hat$** *die wichtiger sind als ich*</NPLT2></ET2>.

Furthermore the tags for the two complex NP spans do not refer to the same representation. One refers to the TH representation the other to the original text. While it is possible to represent this in XML (as multiple trees), it is highly confusing. What is really problematic for an XML representation (or any other inline format) is the addition of 'empty' or 'extra' tokens entered in the target hypothesis, as shown in Table 1. This 'destroys' the token sequence of the original data because the layers are not independent from each other. Further annotation layers (such as competing target hypotheses) can lead to more such interactions.

2.3 Standoff models

As argued above learner corpus architectures should be flexible enough to incorporate additional information without affecting the old data. One reason for that is that otherwise it is impossible to annotate all linguistic layers for all possible target hypotheses (see sentences 1b-g). Another reason is that more than one annotator might want to work on different aspects on the same data. This can only be done if the corpus architecture is flexible enough to allow the following annotation formats.

1. token annotations (annotation values are directly attached to tokens; tokens are technically the smallest unit to be annotated, in many corpora tokens are orthographic words),
2. span annotations (annotation values are attached to a span of consecutive tokens, e.g. topological fields, chunks or any other kind of flat structure which can be expressed as a chain of tokens),
3. tree or graph annotations (hierarchical structures of any kinds; e.g. syntactic structures or discourse structures), and
4. pointing relations (values are attached to elements occurring non-consecutively and widely spread in a text, but do not over each other as in a tree, e.g. anaphoric chains between tokens, spans etc.).

For the remainder of this article we focus on token and span annotation.

In contrast to inline models, standoff models (see e.g. Carletta et al. 2003, Dipper 2005, Chiarcos et al. 2008, Wittenburg 2008, Wörner 2010) separate the original data from the annotations. Each annotation layer is stored

in a separate file; annotations refer to the original data using reference points.[viii] The addition of a new annotation layer is completely independent of the existing layers, as long as the reference is intact. This way it is possible to combine different formats of annotations.

We want to use the second part of the article to demonstrate the need for multiple target hypotheses and a multi-layer standoff architecture using the example of the Falko Essay Corpus.[ix]

3 Case study: Falko

Falko (Lüdeling et al. 2008; Reznicek et al. 2010) is a corpus of written texts by advanced learners of German as foreign language.[x] The learners in the corpus come from different linguistic backgrounds. Data collection is highly controlled and there is a wealth of meta-data for each text which can be used for the creation of ad-hoc subcorpora for specific research questions. The texts in the corpus belong to two writing tasks: summaries and essays. For each task a control corpus of native speaker texts has been compiled under the same conditions. Table 2 shows the corpus size; for the study below we use only the Falko Essays Corpus.

[Table 2]

The learner utterance is pos-tagged and lemmatized using the Tree Tagger (Schmid 1994). Falko can be searched using the multi-layer search tool ANNIS which processes the ANNIS Query Language (Zeldes et al. 2009).xi ANNIS allows a graph-based search across all annotation layers using regular expressions and is thus very powerful.xii

3.1 Target hypotheses in Falko

In the following we want to show in detail how Falko is annotated. We start with a discussion of the target hypotheses. As shown in Section 2.2 the rationale behind a given target hypothesis annotation scheme depends on the research question; and typically an increase of context information leads to a greater distance between the learner text and the TH. The annotation decisions recorded in the guidelines for a specific target hypothesis layer depend therefore directly on how close to the learner data one wants to stay. Two strategies are available:

a. The target hypothesis should stay as close to the learner surface structure as possible.
b. The target hypothesis should reflect as much of the learners' intention in the utterance as possible.

In Falko we formulate two target hypotheses, following these strategies, as exemplified in Table 3. Target hypothesis 1 (TH1), which only corrects clear grammatical errors and orthographic errors, is used for research on morphological and syntactic problems but cannot be used for

research on stylistic errors while target hypothesis 2 (TH2) which is very good for researching lexical problems and stylistic patterns, on the other hand, cannot be used for studying e. g. word order patterns.

[Table 3]

Note that even with very detailed guidelines neither target hypothesis is completely determined. Note also that for specific research questions it might be necessary to add further hypotheses. We will now explain TH1 and TH2 in turn.

3.1.1 Minimal target hypothesis (TH1)

The minimal target hypothesis in the Falko essay corpus consists of a full text that a) differs minimally from the learner text and b) represents a grammatical German sentence at the expense of ignoring errors concerning semantics, pragmatics and style. Where grammar ends and where different levels of correctness apply cannot be solved in general. Nonetheless it is possible to give guidelines so that the decisions for each layer of the corpus are as uniform as possible. In this section we want to illustrate several rules found in the guidelines for each target hypothesis and discuss applications that become possible on the basis of this TH (for the full description see Reznicek et al. 2010).

For all THs changes should be applied to a minimal error exponent, reordering of tokens should span over a minimal amount of tokens and the amount of changes in total should be kept as small as possible, so that the learner structure will stay transparent in all THs to a maximum extent. These general rules need to be specified to deal with specific cases. Let us illustrate this using agreement errors within an NP. In German all elements in an NP need to agree with respect to case, gender, and number. In case of an agreement mismatch within an NP (e.g. a number mismatch between the determiner, an adjective and the head noun), correction will be applied to the adjective(s) first, then to the determiner if necessary. The head noun will be held constant if at all possible. The NP ‘die fleißiege Schüler’ in Table 4 can be corrected in several ways, as illustrated by the options in the last two rows but only one of them is licensed by the rules given above.

[Table 4]

Another example for specific rules concerns word order. In canonical German sentences only one constituent is allowed before the finite verb (see also footnote 5 and Example (2)). However, texts written even by advanced learners of German often show occurrences of two constituents before the finite verb. These errors can be corrected in three ways: move one constituent, move the other constituent, or move the finite verb. To make it easier to search for those sentences with more than one constituent in front of the finite verb we decided to keep the position of the finite verb stable and move its left neighbour constituent to the right, as illustrated in Table 5

[Table 5]

In a similar way the guidelines specify the construction of TH for many possible error situations. Note that this is simply a way of ensuring that similar errors can be found by the same search expression. In no way do we want to imply that we capture any psychological reality.[xiii]
By aligning the target hypothesis with the learner utterance in the manner illustrated above and comparing them we can do a quantitative analysis of underused and overused elements even without any explicit error annotation. Those patterns can be contrasted in turn for learners of different levels of proficiency or L1. A contrastive analysis on the word forms in the Falko essays shows that learners use the reflexive pronoun *sich* significantly less often than the native speakers independently of their L1, while still using it often in total (Zeldes et al. 2008; see Table 6).[xiv] This could be due either to the fact that learners fail to use a reflexive when it is necessary or to the fact that learners simply underuse reflexive verbs. Without a target hypothesis it is impossible to decide between the two options. But doing the same statistics on TH1 reveals that the reflexive is also underused here. From this we can now conclude that learners underuse reflexive verbs.

[Table 6]

Before illustrating how an automatic error analysis can be done on the target hypotheses we want to briefly discuss TH2.

3.1.2 Extended target hypothesis (TH2)
While TH1 concentrates on clear grammatical errors TH2 tries to guess and state the learner's intention. It has often been shown that (even advanced) learners of a foreign language make errors in form-function-mapping (cf. Hendriks 2005; Carroll, Lambert 2006). This is due to often very subtle distribution rules for lexical and structural units; in addition to grammatical rules the learner needs to be aware of register differences, text types, and style. Temporal modification (such as 'in the morning') can be expressed e.g. via an adverb (*morgens*), a prepositional phrase (*am Morgen*), a nominal phrase (*des Morgens*) or in a subordinate clause (*wenn der Morgen anbricht*). None of those alternatives is per se better than any of the others but each of them has its own usage patterns and distribution. It is impossible to understand these patterns or even formalize or code them in an annotation manual. It is immediately obvious that

TH2 is more difficult to construct and keep homogeneous than TH1. One has to keep that in mind when querying the extended target hypothesis.

3.1.3 Word order and information status

With respect to word order TH2 is much freer than TH1. In addition to the clear grammatical rules described above there are ordering patterns that are more difficult to formalize. We want to illustrate this by looking at the middle field (the stretch between the different elements of a verbal complex) in a German sentence. The order of referents in the German middle field is relatively free (Eisenberg 2006). Except for a few cases reordering of constituents does not lead to ungrammatical structures. The order is not arbitrary, however, but serves as a signal for a variety of context sensitive information about the referents such as information structure (Primus 1993; Krifka 2007). [xv] In Table 7 the direct object *einen Arbeit* "a job" has been realized left of temporal adverbial *nach der Universität* "after university". This is a possible word order, but it needs a context which licenses a contrastive reading such as: "after university we try to find a job instead of something else". This reading seems highly improbable in the given context. Therefore the direct object has been placed on the right of the temporal adverbial in the TH2.

[Table 7]

3.1.4 Applications for TH2

TH2 can now be contrasted with TH1 which allows us to retrieve errors concerning semantics, pragmatics as well as problems of register or style. The different patterns in TH2 for learners and native speakers can now serve as a starting point to find candidate structures for semantic, pragmatic and conceptual transfer as well as for fields of L2-specific and universal learning difficulties (Ellis 2009:377). This method is demonstrated in Table 8. The underlined structures mark error regions. The missing definiteness marker in the prepositional phrase *an gesellschaflichen Leben* "in social life" is corrected in both TH1 and TH2. The adverb *gleich* which is ambiguous between "directly" and "equally" is not corrected in TH1 since the "directly" reading leads to a grammatical (albeit probably unintended) sentence. The intention of the adverb is, however, corrected in TH2. In the "equally" reading the structure becomes ungrammatical and so it has been substituted by a different lexeme. Contrasting TH1 with TH2 now filters out grammatical errors (those that are corrected in both THs) and semantic and stylistic errors can be identified.

[Table 8]

3.2 Automatic error tagging

As we have seen, a direct comparison of the learner text with the target hypotheses (and of the target hypotheses with each other) points us to errors on different linguistic levels as long as the levels are aligned with each other. In addition to the qualitative and quantitative comparison of specific structures it is useful to add error annotation. Using automatic

edit tagging, information on differences between two layers (TH1 and LT, for example) can be added in a separate annotation layer. The tag set is given in Table 9.

[Table 9]

The edit tags in Table 9 are similar to the surface error markers (omission, oversuppliance, misformation, misordering etc.) used in (Dulay et al. 1982:150). While relying solely on this error level has been criticized on different occasions (James 2005; Granger 2003) it can be easily automated. Used in combination with the target hypotheses it offers a rich way of filtering query results for CIA and EA.

In order to illustrate this let us come back to the example of multiple constituents before the finite verb in German (Example (2), Table 10). Without further manual annotation and only based on edit tags and the target hypotheses it now becomes possible to answer the following research question: How often do we find multiple constituents before the finite verb in learners and in native speakers? Using the edit tags we can formulate a search for tokens that occur between a token tagged as end of a sentence on the left and a finite verb on the right that is tagged as "MOVS" for the TH1. We can formulate an additional restriction that there must be further tokens between the finite verb and the end of the sentence to the right.[xvi] We can then see that there is no error of this type in the L1 corpus while there are 20 errors of this type in the learner data.

Since THs are full text layers we can add any other kind of annotation, such as POS or lemma annotation. This means that queries can be made even more specific, see Table 10.

[TABLE 10]

POS annotation becomes even more interesting if one seeks to find deviations on POS tags and POS chains directly (Aarts, Granger 1998; Borin, Prütz 2004; Zeldes et al. 2008). Once again this information can be incorporated into the corpus, this time by using edit tags for differences in the POS annotation layers for LT and the THs. The same holds for the lemmas.

3.3 Manual error tagging

While automatic edit tags might be useful the objective of many learner corpus studies is a more fine-grained and linguistically informed error classification. This has been done in the Falko Essay Corpus for all complex verbs. Again the layered representation allows splitting the annotations into different classes: verb category, verb lemma, verb error type, and verb form. Those can then be recombined again for specific queries.

Table 11 shows a sentence in the Falko essay learner corpus with all annotations.[xvii]

[TABLE 11]

4 Summary

In this chapter we have shown, why the question of corpus architecture matters. We argued for a multi-layer standoff architecture at least for small specialised corpora like the learner corpus Falko for the following reasons: Independent annotation layers allow a wide range of structurally different annotation types, they prevent spreading of errors, and they ensure the readability of all annotation layers independent of their number and the sustainability of the data storage. All layers can then be recombined ad-hoc in query processors like ANNIS. We have demonstrated why competing explicit target hypotheses are necessary to allow a well-documented error analysis on very different linguistic levels. Including those target hypotheses directly into the corpus allows for a list of automatically derived data enhancements like surface edit tags to be generated which allow very specific queries on higher levels of abstraction like POS or lemma sequences and their deviations on different THs without further manual annotation.

5 Bibliography

Aarts, J. & Granger, S. 1998. Tag Sequences in Learner Corpora: A key to interlanguage grammar and discourse. In *Learner English on computer*. S. Granger (ed.), 132–141. London: Longman.

Abe, M. 2004. A Corpus-based Analysis of Interlanguage: Errors and English proficiency Level of Japanese Learners of English. In *Handbook of an International Symposium on Learner Corpora in Asia (ISLCA)*, 28–32.

Belz, J.A. 2004. Learner Corpus Analysis and the Development of Foreign Language Proficiency. *System* 32: 577–591.

Bird, S. & Liberman, M. 2001. A Formal Framework for Linguistic Annotation. *Speech Communication* 33: 23–60.

Borin, L. & Prütz, K. 2004. New Wine in old Skins?: A Corpus Investigation of L1 Syntactic Transfer in Learner Language. In *Corpora and Language Learners.* G. Aston & S. Bernardini & D. Stewart (eds), 67–87. Amsterdam, Philadelphia: John Benjamins.

Boyd, A. 2010. EAGLE: An Error-Annotated Corpus of Beginning Learner German. In *Proceedings of the LREC.* Valletta, Malta.

Breckle, M. & Zinsmeister, H. 2010. Zur lernersprachlichen Generierung referierender Ausdrücke in argumentativen Texten. In *Textmuster: schulisch - universitär - kulturkontrastiv.* D. Skiba (ed.), 79–101. Frankfurt a. M.: Peter Lang.

Carletta, J. & Evert, S. & Heid, U. & Kilgour, J.R. & Voormann, H. 2003. The NITE XML Toolkit: Flexible Annotation for Multimodal Language Data. *Behavior Research Methods, Instruments, and Computers* 35: 353–363.

Carroll, M. & Lambert, M. 2006. Reorganizing Principles of Information Structure in Advanced L2s: French and German Learners of English. In *Educating for Advanced Foreign Language Capacities. Constructs, Curriculum, Instruction, Assessment.* H. Byrnes & H. Weger-Guntharp & K.A. Sprang (eds), 54–73. Washington, DC.

Chiarcos, C. & Dipper, S. & Götze M. & Ritz, J. & Stede, M. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In *Proceeding of the Conference on Global Interoperability for Language Resources, Hong Kong, January 2008*.

Chuang, F.-Y. & Nesi, H. 2006. An Analysis of Formal Errors in a Corpus of L2 English produced by Chinese Students. *Corpora* 1: 251–271.

Crompton, P. 2005. 'Where', 'In Which', and 'In That': A Corpus-Based Approach to Error Analysis. RELC Journal 36: 157–176.

Dagneaux, E.& Denness, S. & Granger, S. & Meunier, F. 1996. *Error Tagging Manual Version 1.1*. Louvain-la-Neuve: Université catholique de Louvain. Centre for English Corpus Linguistics.

Dagneaux, E. & Denness, S. & Granger, S. 1998. Computer-aided Error Analysis. *System* 26: 163–174.

Díaz-Negrillo, A. & Fernández-Domínguez, J. 2006. Error Tagging Systems for Learner Corpora. *Revista Española de Lingüística Aplicada* 19: 83–102. Available online at http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810.

Dipper, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of Berliner XML Tage (BXML 2005)*, 39–50. Berlin.

Dulay, H. & Burt, M.; Krashen, S. 1982. *Language Two*. New York, Oxford: Oxford University Press.

Eisenberg, P. 2006. *Der Satz*. 3rd ed. Stuttgart: Metzler.

Ellis, R. 2009. *The Study of Second Language Acquisition.* New York, Oxford: Oxford University Press.

Fitzpatrick, E. & Seegmiller, S.M. 2001. The Montclair electronic language learner database. In *Proceedings of the International Conference on Computing and Information Technologies.* G. Antoniou & D. Deremer (eds). World Scientific.

Fitzpatrick, E. & Seegmiller, S.M. 2004. The Montclair electronic language database project. In *Applied Corpus Linguistics: A Multidimensional Perspective.* U. Connor & T.A. Upton (eds).Amsterdam,New York: Rodopi.

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

Granger, S. 2008. Learner corpora. In *Corpus linguistics: An international Handbook.* A. Lüdeling & M. Kytö (eds), 259–275. Berlin, New York: Mouton de Gruyter.

Granger, S. & Dagneaux, E. & Meunier, F. & Paquot, M. 2009. *The International Corpus of Learner English. Version 2*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Hendriks, H., ed. 2005. *The Structure of Learner Varieties*. Berlin, New York: Mouton de Gruyter.

Höhle, T.N. 1986. Der Begriff 'Mittelfeld': Anmerkungen über die Theorie der topologischen Felder. In *Kontroversen, alte und neue: Akten des VII. Kongresses der Internationalen Vereinigung für germanische Sprach- und Literaturwissenschaft.* A. Schöne & I. Stephan (eds), 329–340. Tübingen: Niemeyer.

Izumi, E. & Uchimoto, K. & Isahara, H. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management* 12: 119–125.

James, C. 2005. *Errors in Language Learning and Use: Exploring Error Analysis.* Repr. [Applied linguistics and language study]. London: Longman.

King, P.R. & Munson, E.V. (eds). 2004. DDEP-PODDP 2000. Berlin: Springer.

Krifka, M. 2007. Basic Notions of Information Structure. In *Interdisciplinary Studies of Information Structure 6.* C. Fery & M. Krifka (eds). Potsdam.

Lehmberg, T. & Wörner, K. 2008. Annotation standards: 22. In *Corpus linguistics: An international Handbook.* A. Lüdeling & M. Kytö (eds), 484–501. Berlin, New York: Mouton de Gruyter.

Lenerz, J. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. München, Tübingen: Narr.

Lennon, P. 1991. Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics* 12: 180–196. Available online at http://applij.oxfordjournals.org/cgi/content/abstract/12/2/180.

Lüdeling, A. 2007. Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In *Sprachkorpora - Datenmengen und Erkenntnisfortschritt.* W. Kallmeyer & G. Zifonun (eds), 28–48. Berlin, New York: Mouton de Gruyter.

Lüdeling, A. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung.* M. Walter & P. Grommes (eds), 119–140. Tübingen: Max Niemeyer Verlag.

Lüdeling, A. & Doolittle, S. & Hirschmann, H. & Schmidt, K. & Walter, M. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 45: 67–73.

Lüdeling, A. to appear. Corpora in Linguistics: Sampling and Annotation. In *Going Digital: Evolutionary and Revolutionary Aspects of Digitization.* K. Grandin (ed.). USA, New York: Science History Publications.

Lüdeling, A: & Hirschmann, H. & Rehbein, I. & Reznicek, M. & Zeldes, A. 2010. *Syntactic Overuse and Underuse: A Study of the Parsed Learner Corpus Falko*. Presentation given at the 9th Treebanks and Linguistic Theory Workshop, Tartu, December 2010.

Primus, B. 1993. Word Order and Information Structure: A Performance Based Account of Topic Positions and Focus Positions. In *Syntax.* J. Jacobs & A.v. Stechow & W. Sternefeld & T. Vennemann (eds), 880–896. Berlin, New York: Mouton de Gruyter.

Reznicek, M.& Walter, M. & Schmidt, K. & Lüdeling, A. & Hirschmann, H.; Krummes, C. & Andreas, T. 2010. *Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 1.0*. Berlin: Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin Available online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. Available online at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics* 10: 209–231.

Sperberg-McQueen, C. 1999. Concurrent document hierarchies in MECS and SGML. *Literary and Linguistic Computing* 14: 29–42.

Tenfjord, K. & Hagen, J.E. & Johansen, H. 2006. The «Hows» and the «Whys» of Coding Categories in a Learner Corpus: or «How and Why an Error-Tagged Learner Corpus is not 'ipso facto' One Big Comparative Fallacy ». *Rivista di psicolinguistica applicata*: 93–108.

Tono, Y. 2004. Multiple Comparisons of IL, L1 and TL Corpora: The Case of L2 Acquisition of Verb Subcategorization Patterns by Japanese Learners of English. In *Corpora and Language Learners.* G. Aston & S. Bernardini & D. Stewart (eds), 45–66. Amsterdam, Philadelphia: John Benjamins.

van Rooy, B. & Schäfer, L. 2002. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics & Applied Language Studies* 20: 325.

Weinberger, U. 2002. Error analysis with computer learner corpora: A corpus-based study of errors in the written German of British University Students. MA thesis. Lancaster: Lancaster University

Wittenburg, P. 2008. Preprocessing Multimodal Corpora. In *Corpus Linguistics: An International Handbook.* A. Lüdeling & M. Kytö (eds), 664–685. Berlin, New York: Mouton de Gruyter.

Wörner, K. 2010. A Tool for Feature-Structure Stand-Off-Annotation on Transcriptions of Spoken Discourse. In *Proceedings of the Seventh conference on International Language Resources and Evaluation: LREC'10.* N. Calzolari & K. Choukri & B. Maegaard & J. Mariani & J. Odijk & S. Piperidis & M. Rosner & D. Tapias (eds). Valletta, Malta: European Language Resources Association (ELRA). Available online at http://www.exmaralda.org/files/LREC_Sextant.pdf.

Wörner, K. & Witt, A. & Rehm, G. & Dipper, S. 2006. Modelling Linguistic Data Structures. In *Proceedings of Extreme Markup Languages.* Montreal.

Zeldes, A. & Lüdeling, A. & Hirschmann, H. 2008. What's hard?: Quantitative evidence for difficult constructions in German learner data. In *Proceedings of QITL 3.* Helsinki. Available online at http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Presentations/Zeldes_et_al.ppt

Zeldes, A. & Ritz, J. & Lüdeling, A. & Chiarcos, C. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.*

Zinsmeister, H. & Breckle, M. 2010. Starting a sentence in L2 German: Discourse annotation of a learner corpus. In *Semantic approaches in natural language processing: Proceedings of the Conference on Natural Language Processing 2010.* M. Pinkal (ed.), 181–185. Saarbrücken: Universaar.

All URLs were checked on 12/10/2010.

LT	weil	er	die	Ziele		,	die	wichtiger	als	ich	sind	,	hat			.
	because	*he*	*the*	*goals*		,	*that*	*more-important*	*than*	*I*	*are*	,	*has*			.
NP			NP													
TH	weil	er		Ziele	hat	,	die	wichtiger			sind			als	ich	.
	because	*he*		*goals*	*has*	,	*that*	*more-important*			*are*			*than*	*I*	.
NP				NP												

Table 1: Competing and overlapping annotation spans for complex noun phrases for the learner text (LT) and the target hypothesis (TH)

Falko (texts/ tokens)	Essays	Summaries
Learner texts (L2)	248/ 122789	107/ 40787
Native speaker control group (L1)	95/ 68485	57/ 21184

Table 2: Texts and tokens in Falko

form → function

Minimal target hypothesis (TH1)	**Extended target hypothesis (TH2)**
minimal grammatical corrections, sentence-based	recourse to semantic and pragmatic information, text-based

→TH is grammatically correct	→TH is grammatically correct, semantically coherent and pragmatically acceptable
+ relatively clear-cut annotation guidelines + high inter-annotator accuracy possible + structural proximity to the learner utterance	+ intended proximity to the learner's intention + inclusion of 'higher-level' lin guistic information
- may still contain errors	- is open to more varied interpretations - may lead to substantial changes in the surface structure

Table 3: TH1 and TH2 in the Falko corpus

LT	dadurch	kann	man	die	fleißiege	Schüler	schaffen
	thus	*can*	*one*	*the*	*diligent*	*students*	*produce*
	"in this way diligent students can be produced"						
TH1	dadurch	kann	man	die	fleißigen	Schüler	schaffen
!TH1	dadurch	kann	man		fleißige	Schüler	schaffen

Table 4: Illustration of TH1 for agreement errors in a learner utterance (FalkoEssayL2v2_0:usb012_2006_10). !TH1 is a grammatically possible target hypothesis which is rejected by the guidelines.

LT	Und	dann		jede	bekommt		eine	finanzielle	Entlohnung	.
	and	*then*		*everyone*	*receives*		*a*	*financial*	*reward*	.
TH1	Und	dann			bekommt	jede	eine	finanzielle	Entlohnung	.
!TH1	Und	dann	bekommt	jede			eine	finanzielle	Entlohnung	.

Table 5: Illustration of word order errors in TH1 of a learner utterance (FalkoEssayL2v2_0:fkb015_2008_07). !TH1 is a grammatically possible target hypothesis which is rejected by the guidelines.

lemma	de	da	en	fr	pl
in	0.012261	0.014041	0.014247	0.015272	0.012135
es	0.011945	0.0109	0.011379	0.013347	0.008163
sie	0.008193	0.010643	0.008835	0.010909	0.006067
man	0.0079	0.012438	0.008742	0.009754	0.00695
dass	0.007404	0.012823	0.008789	0.009625	0.00888
von	0.007122	0.007309	0.006846	0.007315	0.010259
auch	0.008362	0.008527	0.005828	0.005775	0.005461
für	0.007201	0.006091	0.007216	0.006802	0.005736
sind	0.004271	0.008976	0.007308	0.00693	0.004964
sich	**0.0117**	**0.006283**	**0.006291**	**0.00693**	**0.00717**

ich	0.003877	0.013272	0.005366	0.003465	0.001434
aber	0.003347	0.007309	0.006245	0.007315	0.003365

Table 6: Overuse/ underuse visualization on word forms in Falko original data. The frequencies of each lemma in the L1 data (column 'de') are compared with the frequencies in different L2 groups (the column titles give their native languages: da-Danish, en-English, fr-French, pl-Polish, ru-Russian). Plain numbers signal overuse, underlined ones signal underuse; the darker the cell the stronger the overuse or underuse (Zeldes et al. 2008).

LT

Wenn wir Universitätsprüfung bestehen, haben wir sehr Glück nach anderen Menschen. Denn wir hoffen, dass wir [**einen Arbeit**] [nach der Universität] finden.
*If we University-exam pass, have we a-lot-of luck after other people. Because we hope that we [**a job**] [after the university] find.*

TH2

Wenn wir eine Universitätsprüfung bestehen, haben wir der Meinung anderer Menschen nach viel Glück. Denn wir hoffen, dass wir [nach der Universität] [**eine Arbeit**] finden.
*If we a university-exam pass have we the opinion of-other people after a-lot-of luck. Because we hope that we [after the university] [**a job**] find.*

There are people who think that we are quite lucky if we

pass the university exam. Because we hope to find a job after university.

Table 7: Falko example (LT) plus target hypothesis 2 (TH2) for FalkoEssayL2v2.0:trk006_2006_05. TH2 here corrects the word order in the middle field.

LT	Die Frauen hatten den Wunsch, an gesellschaflichen Leben teilzunehmen und gleich wie Männer zu arbeiten. *The women had the wish, on social life to-take-part and directly/equally like men to work.*
TH1	Die Frauen hatten den Wunsch, **am** gesellschaftlichen Leben teilzunehmen gleich wie Männer zu arbeiten. *The women had the wish, **on-the** social life to-take-part and directly like men to work.*
TH2	Die Frauen hatten den Wunsch, **am** gesellschaftlichen Leben teilzunehmen und **genauso** wie **die** Männer **arbeiten zu gehen.** *The women had the wish, on-the social life to-take-part and **equally like** men **to work**.*

Table 8: Falko example (LT) and two target hypotheses (TH1, TH2) for FalkoEssayL2v2.0:fk019_2006_07. The target hypotheses can be contrasted to find higher-level errors such as wrong lexical choice for the ambiguous word *gleich* standing for “immediately” and “equally”.

Tag	Description
INS	inserted token in TH
DEL	deleted token in TH
CHA	changed token in TH
MOVS	source location of moved token in TH
MOVT	target location of moved token in TH
MERGE	tokens merged in TH
SPLIT	tokens splitted in TH

Table 9: Surface deviance “edit tags” used in the Falko essay corpus

LT	In	diesem	Fall	auf	solche	Leute	können	die	Freunden				wirken	.
	In	*this*	*case*	*on*	*those*	*people*	*can*	*the*	*friends*				*have-an-impact*	.
pos	APPR	PDAT	NN	APPR	PIAT	NN	VMFIN	ART	NN				VVINF	$.
Lemma	in	dies	Fall	auf	solch	Leute	können	d	Freund				wirken	.
TH1	In	diesem	Fall				können	die	Freunde	auf	solche	Leute	wirken	.

TH1pos	APPR	PDAT	NN				VMFIN	ART	NN	APPR	PIAT	NN	VVINF	$.
TH1lemma	in	dies	Fall				können	d	Freund	auf	solch	Leute	wirken	.
TH1Diff				**MOVS**	**MOVS**	**MOVS**			CHA	**MOVT**	**MOVT**	**MOVT**		
TH2	In	diesem	Fall	auf	solche	Leute	können	die	Freunde	auf	solche	Leute	einwirken	.
TH2pos	APPR	PDAT	NN				VMFIN	ART	NN	APPR	PIAT	NN	VVINF	$.
TH2lemma	in	dies	Fall				können	d	Freund	auf	solch	Leute	einwirken	.
TH2Diff				**MOVS**	**MOVS**	**MOVS**			CHA	**MOVT**	**MOVT**	**MOVT**	CHA	

Table 10: Learner utterance (LT) plus target hypotheses (TH1, TH2) and error tags for FalkoEssayL2v2.0:usb008_2006_10. Each layer is automatically pos-tagged and lemmatized. Edit tags like “MOVS” help find word order errors in the target hypotheses.

LT	word	darüber	negativ	ausgesprochen,		dass	sie		mit	dem	Firmen	mehr	direkt	arbeiten
		over.it negatively spoken.out that they with the.SG enterprises.PL more direct work.3.PERS.PL												
auto annotation	pos	PROAV	ADJD	VVPP		KOUS	PPER		APPR	ART	NN	ADV	ADJD	VVFIN
	lemma	darüber	negativ	aussprechen		dass	sie		mit	d	Firma	mehr	direkt	arbeiten
minimal target hypothesis	TH1	dazu	negativ	ausgesprochen	,	dass	sie		mit	den	Firmen	direkter		arbeiten
	TH1pos	PROAV	ADJD	VVPP	$,	KOUS	PPER		APPR	ART	NN	ADJD		VVFIN
	TH1posDiff											MERGE		
	TH1lemma	dazu	negativ	aussprechen	,	dass	sie		mit	d	Firma	direkt		arbeiten
	TH1lemmaDiff	CHA			INS							MERGE		
	TH1Diff	CHA			INS					CHA		MERGE		
extended target hypothesis	TH2	dazu	negativ	ausgesprochen	,	um		direkter	mit	den	Firmen			zusammenzuarbeiten
	TH2pos	PROAV	ADJD	VVPP	$,	KOUI		ADJD	APPR	ART	NN			VVINF
	TH2posDiff				INS	CHA	DEL	MOVT				MOVS	MOVS	CHA
	TH2lemma	dazu	negativ	aussprechen	,	um		direkt	mit	d	Firma			zusammen-arbeiten
	TH2lemmaDiff	CHA			INS	CHA	DEL	MOVT				MOVS	MOVS	CHA
	TH2Diff	CHA			INS	CHA	DEL	MOVT		CHA		MOVS	MOVS	CHA

Complex verb target hypothesis	THverb	dazu	negativ	geäußert	,	um		direkter	mit	den	Firmen			zusammenzuarbeiten
	THverbpos	PROAV	ADJD	VVFIN	$,	KOUI		ADJD	APPR	ART	NN			VVINF
	THverblemma	dazu	negativ	geäußert	,	um		direkt	mit	d	Firma			zusammenarbeiten
	THverbDiff	CHA		CHA	INS	CHA	DEL	MOVT		CHA		MOVS	MOVS	CHA
Complex verbs error tags	verbkategorie			vpart										
	verblemma			aussprechen										
	verbfehlertyp			sem										
	verbform			p2										

Table 11: Fragment of a learner utterance FalkoEssayL2v2.0:fk001_2006_08: [Aus diesem Grund haben sich die Universitäten] darüber negativ ausgesprochen, dass sie mit den Firmen mehr direkt arbeiten“, roughly [for that reason the universities]“ spoke negatively about the fact that they wanted to work more closely with the companies”) with annotations for three target hypotheses and error annotation on the complex verbs.

Notes:

[i] One interesting exception is the Montclair electronic learner database (Fitzpatrick, Seegmiller 2001, 2004) which limits itself to a target hypothesis.

[ii] There has been a long and controversial discussion about the concept of an 'error' language acquisition research. We will not discuss this here due to space constraints but see Lennon (1991); Ellis (2009)

iii http://www.latl.unige.ch/ [checked 06/12/2010].

[iv] The sentence is a translation of the German learner utterance from FalkoEssayL2v2_0:fk012_2006_07 (for references to the corpus see Section 3).

[v] See the topological model for German sentences (Drach 1937; Höhle 1986).

[vi] XML formats are much more sustainable than proprietary formats, especially if they adhere to one of the accepted standards like TEI (Lehmberg, Wörner 2008). Note that we do not argue against XML here, only against XML inline formats. We also use an XML format to store our data; see below.

[vii] There are, of course, ways of dealing with overlapping spans in XML (for an overview see Sperberg-McQueen 1999 and King, Munson 2004).

[viii] Since standoff models were originally developed for multimodal corpora the reference is often coded with regard to a 'timeline' (taken from the audio or video layer, cf. Bird, Liberman 2001; Carletta et al. 2003). In multi-layer corpora that have no timeline the token sequence is used as the reference (Wörner et al. 2006; Wittenburg 2008).

[ix] Falko was, to our knowledge, the first learner corpus with a multi-layer standoff architecture. Other learner corpora such as EAGLE (Boyd 2010) and Alesko (Breckle, Zinsmeister 2010; Zinsmeister, Breckle 2010) are now also based on this architecture.

[x] The corpus with the target hypotheses and all annotations is freely available at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/-forschung-en/falko/standardseite-en.

[xi] The tool is freely available at http://www.sfb632.uni-potsdam.de/d1/annis/.

[xii] Technically ANNIS operates on a relational database. In addition it is stored in a sustainable XML format (PAULA-XML; Dipper 2005, Chiarcos et al. 2008) and relAnnis (Zeldes et al. 2009).

[xiii] Just as an aside: Even if at first sight it seems counterintuitive, it is necessary to construct a target hypothesis for our native speaker control groups as well.

The comparison of the differences in the overuse/underuse patterns for L1 and L2 then shows what kind of errors occur more often in learner texts only. A surprising finding is that L1-texts contain significantly more punctuation errors than the learner texts.

[xiv] Zeldes et al. (2008) as well as Lüdeling (to appear) argue that underuse might be a diagnostics for learning difficulty. For reasons of space we cannot pursue this here; nor can we go into the reasons for the underuse of the reflexive.

[xv] In addition word order preferences in the middle field are constrained by other factors such as phonological weight, grammatical function, animacy of the referent etc. (Lenerz 1977).

[xvi] Sentences consisting only of prefield and finite verb like *er schläft* "he sleeps" are not found that way, but those can easily be found in a second query. The search can be easily formulated in ANNIS. ANNIS has a permanent URL for result sets – you can see the query and the results at http://korpling.german.hu-berlin.de/falko-suche/Cite/AQL(pos%3D%20%22%24.%22%26%0AZH1Diff%3D%22MOVS%22%26%0AZH1pos%3D/V.FIN/%26%0Apos%3D/V.+/%26%0A%231.1%2C6%232%26%0A%232.%233%26%0A%233_%3D_%234),CIDS(FalkoEssayL2V2_0),CLEFT(10),CRIGHT(10)

[xvii] In addition to the token-based annotation and span-annotation the learner utterances and the THS are also automatically parsed (using the Berkeley parser, trained on the TiGer Treebank, Lüdeling et al. 2010).