Multilevel Learner Corpora

This paper presents multilevel corpus architectures as a means of studying learner language. Limitations of current inline-annotated corpora are discussed and contrasted with advantages offered by state-of-the-art rich annotation architectures. An example using the Annis web interface shows first steps in applying such architectures to learner data.

Corpus-based Analysis of Learner Language

Empirical data plays an important role in studying learner language, and corpora are a major resource for such studies. They are used for (1) systematic error analysis (EA) as well as (2) examining learner interlanguage and its over-/underuse of categories wrt. native usage, working from the assumption that learner language is an independent variety with its own linguistic system (Selinker 1972; for an overview of corpus studies see e.g. Granger et al. 2002, Granger 2008). For purpose (1), errors and other categories are annotated, usually using familiar inline architectures where annotations appear alongside raw data. Since these annotations may apply to multiple tokens, corpus designers tend to use XML or other tags and headers to mark up the relevant data. The International Corpus of Learner English (ICLE, Granger 1993), for example, marks errors as in: *the poet wants to <?jake?>* (ICLE-IT-BER-0001.10), and the ICLE guidelines for error classification (Dagneaux et al. 2005, 13) offer error types and a target hypothesis, e.g.: *despite (FS) it's its efforts* (FS stands for the error *Form, Spelling*).

Purpose (2) is especially connected with Contrastive Interlanguage Analysis (CIA) (Granger et al. 2002, 12f), so called since learner data is contrasted with comparable native or other learner data. The contrast may be drawn qualitatively or quantitatively, by either examining surface phenomena (word forms) or categorizing them with annotations.

In the following section we argue that for both purposes (1) and (2) an inline corpus architecture is often insufficient: This type of architecture can neither deal with conflicting annotation spans, nor with multiple or alternative values within one annotation level. Therefore, we suggest the use of multi-layer architectures for learner corpora and demonstrate how such an architecture can be applied to learner data in learner corpus studies.

Multilevel Corpora and Applications for Learner Language

The limitations of single-layer architectures become clear when annotation complexity grows. This is true for error annotation as well as for the description of canonical learner utterances, which is necessary for an adequate analysis of learner interlanguage.

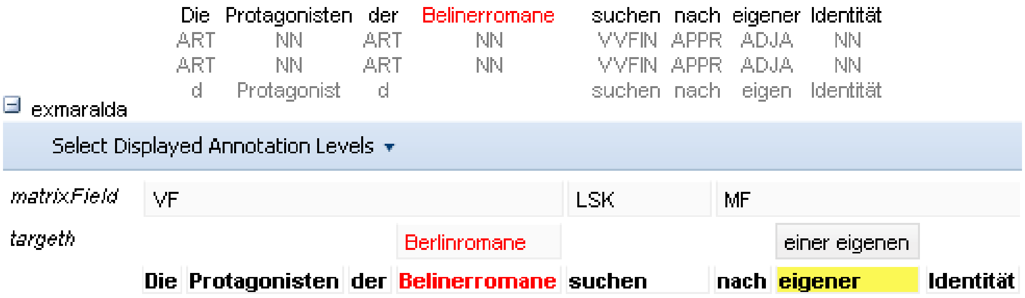
Not only can annotations be recursively embedded (e.g. spelling errors within a syntactic error), but conflicting hierarchies may defy inline representation. For example, annotating prepositional objects and argument structure errors produces invalid XML in the following sentence: *He <argerr> awaited <ppobj> for </argerr> his wife</ppobj>*. We find that either the preposition or the verb is erroneous (*waited for* or *awaited* are possible hypotheses), yet the prepositional object to be tagged already begins within this error span, continuing until *wife*. This sentence shows that any error analysis is based on an (implicit or explicit) target hypothesis. Since the uncertainty regarding the correct target hypothesis may be high, multiple, competing annotations are in fact necessary (cf. Lüdeling 2008 on the very low inter-annotator agreement for target hypotheses and the consequences for analyses based on error annotation). Integrating target hypotheses for the above sentence also illustrates the second main deficit of single-level inline architecture: Since there is only one hypothesis level we can only choose one alternative, though in fact many may be possible.

Most CIA studies compare the frequencies of word forms. For this purpose no annotation of the learner data is necessary. Many questions in language acquisition research, however, require the comparison of categories other than word forms, like part-of-speech categories, information structure categories, morphological or syntactic categories etc.

The solution offered by multi-layer architectures is to separate the data from the annotations. Each annotation level can appear in a separate file pointing to elements in the raw data, which also allows the addition of new annotation levels as they become necessary, without affecting existing data. Multiple annotations on the same level may also refer to the same elements, allowing for example several target hypotheses for the same error. This is achieved using a stand-off XML format, in which annotation elements refer to data using unique identifiers (see e.g. Carletta et al. 2003). An example of an architecture utilizing such a format is Annis2, using PAULA XML (Dipper 2005).

Annis2 – A Multilevel Corpus Architecture

Annis2 is a web-based corpus interface built to query and visualize multilevel corpora. It can formulate exact and pattern queries on arbitrary, possibly nested annotation levels, which may be conflictingly overlapping, discontinuous or carry multiple values for the same annotation level in the same corpus position. This allows representing annotations applying to spans of tokens and hierarchical trees (e.g. syntax), among other things.



Die	Protagonisten	der	Belinerromane	suchen	nach	eigener	Identität
ART	NN	ART	NN	VVFIN	APPR	ADJA	NN
ART	NN	ART	NN	VVFIN	APPR	ADJA	NN
d	Protagonist	d		suchen	nach	eigen	Identität

exmaralda

Select Displayed Annotation Levels ▾

matrixField	VF				LSK	MF		
targeth				Berlinromane			einer eigenen	
	Die	Protagonisten	der	Belinerromane	suchen	nach	eigener	Identität

Fig. 1: Screenshot of the Falko German learner corpus (Lüdeling et al. 2008) in Annis2. The sentence translates to “The protagonists of the Berlin novels search for their own identity.” The errors are a word formation error in “Berlinerromane” and a missing article in “eigener”.

By allowing queries on multiple, conflicting annotation levels simultaneously, the system enables the study of interdependencies between a potentially limitless variety of annotation levels. We suggest this architecture is optimally suited to bringing subtle annotation levels already found in L1 corpora to the study of learner interlanguage. In this context we have carried out pilot studies using the German learner corpus Falko for both EA and CIA purposes.

References

Carletta, J./Evert, S./Heid, U./Kilgour, J./Robertson, J./Voormann, H. (2003) The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data. *Behavior Research Methods, Instruments, and Computers* 35(3), 353-363.

Dagneaux, E./Denness, S./Granger, S./Meunier, F./Neff, J./Thewissen, J. (2005), *Error Tagging Manual. Version 1.2.* Louvain: Université catholique de Louvain, Centre for English Corpus Linguistics.

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

Granger, S. (1993) The International Corpus of Learner English. In: Aarts J./de Haan P./Oostdijk N. (eds.), *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, 57-69.

Granger, S. (2008), Learner Corpora. In: Lüdeling, A./Kytö, M. (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 259-275.

Granger, S./Hung, J./Petch-Tyson, S. (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

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

Lüdeling, A./Doolittle, S./Hirschmann, H./Schmidt, K./Walter, M. (2008) Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2/2008, 67-73.

Selinker, L. (1972), Interlanguage. *International Review of Applied Linguistics* 10, 201-231.