RESEARCH NOTE

The *Kansas Developmental Learner corpus* (KANDEL): A developmental corpus of learner German

Nina Vyatkina

The University of Kansas

This article presents the Kansas Developmental Learner corpus (KANDEL), a corpus of L2 German writing samples produced by several cohorts of North American university students over four semesters of instructed language study. This corpus expands the number of freely and publicly available learner corpora while adding to the depth of these corpora with a unique set of features. It does so by focusing on an L2 other than English, German, targeting beginning to intermediate L2 proficiency levels, and including dense developmental data and annotations for multiple linguistic variables, learner errors, and over twenty learner and task variables. Furthermore, this article reports the procedure and results of an inter-annotator agreement study as well as an in-depth analysis of annotator disagreement. In this way, it contributes to best practices of annotating learner corpora by making the annotation process transparent and demonstrating its reliability.

Keywords: longitudinal learner corpus; beginning and intermediate L2 proficiency; written corpus; L2 German; error annotation; inter-annotator agreement

1. Introduction

This article presents the *Kansas Developmental Learner corpus* (KANDEL), a corpus compiled from L2 German writing samples produced by North American university students over four semesters of instructed language study. This corpus was collected, annotated, and analyzed as part of a project on Instructed Second Language Acquisition (ISLA) that investigated learner linguistic development from beginning to intermediate L2 proficiency levels.

The burgeoning field of Learner Corpus Research (LCR) has provided ISLA scholars with resources and tools for testing hypotheses on L2 development on the basis of data from learner corpora (for overviews and discussion, see Granger 2015; Granger et al. 2015; Tono as cited in Callies & Paquot 2015). However, both fields have a number of common gaps repeatedly pointed out by prominent scholars, notably the preponderance of cross-sectional over longitudinal studies. Especially rare are both ISLA studies (see Larsen-Freeman 2006; Ortega & Byrnes 2008) and LCR studies (see Alexopoulou et al. 2015; Gries & Deshors 2015) that explore dense developmental data, data collected from learners at beginning levels, and multidimensional studies that integrate various types of data and metadata. Finally, the overwhelming majority of available learner corpora are limited to L2 English, although the number of corpora containing L2 data from other languages has recently been growing. Many of these corpora are listed on the 'Learner Corpora around the World' webpage maintained by the *Centre for English Corpus Linguistics* (CECL) at Louvain-la-Neuve, Belgium.[1]

[1] https://www.uclouvain.be/en-cecl-lcworld.html (accessed 4 March 2016).

This project aims to address these gaps. First, KANDEL is a developmental learner corpus that is not only longitudinal but also dense, as learner writing samples were collected every three to five weeks over several academic semesters. In this way, KANDEL is different from not only cross-sectional corpora but also from the few available longitudinal corpora: the Georgetown corpus (Byrnes et al. 2010; Lüdeling et al. 2005) and the *Longitudinal Database of Learner English* (LONGDALE, Meunier & Littré 2013), the data for which were collected once per year or curricular level (but see Ott et al. 2012 for a description of the *Corpus of Reading Comprehension Exercises in German* (CREG), KANDEL's twin developmental corpus). Second, KANDEL participants are *ab initio* learners, whereas the overwhelming majority of learner corpora comprise data from (high) intermediate to advanced learners. Rare exceptions are the *International Corpus of Crosslinguistic Interlanguage* (ICCI) and the *National Institute of Communications Technology Japanese Learner English Corpus* (NICT JLE; see Tono as cited in Callies and Paquot 2015) as well as the *Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context* (MERLIN; Wisniewski et al. 2013). This feature of KANDEL is especially significant because the absence of language data from beginning proficiency levels has so far prevented explorations of the full course of instructed SLA (Ortega & Byrnes 2008; Ortega & Sinicrope 2008). Third, following the best practices outlined by the Falko corpus team (e.g. Reznicek et al. 2013), KANDEL writing samples have been annotated for multiple linguistic features and learner errors, and cross-tabulated with over twenty learner and task variables, which allows for multifactorial analyses. Fourth, KANDEL complements the small but growing body of corpora representing L2 German (for an overview, see Krummes & Ensslin 2014) by adding developmental data from beginning writers. Last but not least, the corpus is freely and publicly available for online searches and for download using the Creative Commons license.[2]

In what follows, I describe the corpus collection process, the structure of its data and metadata, and the data annotation procedures including a report of inter-annotator agreement and an in-depth analysis of annotator disagreement. I conclude by summarizing the results of the first KANDEL-based studies and suggestions for future research.

2. Description of the corpus

2.1 Participants

The data were collected from several cohorts of students who enrolled in a basic German language program at a large public US-American university. Students in this program are typically 18-23 years old, with an approximately equal number of females and males. An overwhelming majority of the students grew up in the American Midwest, have American English as their L1, and have little exposure to German outside the classroom. Although a portion of this learner population is of German descent, very few students have relatives who speak German. Also, virtually none of these students have lived in German-speaking countries for an extended period of time, although some have travelled there for a few weeks (either for school trips or on vacation). Therefore, their learner language can be described as a prototypical German as a Foreign Language variety. The L2 German proficiency level of these learners is at

[2] https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/research/kandel (accessed 4 March 2016).

or below the A2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001), as reported in Vyatkina (2016). Furthermore, they have either no knowledge or a low level of knowledge of additional languages (most commonly Spanish).

To summarize, this learner population is fairly homogenous vis-à-vis their ethnographic and language knowledge background. However, there is still variation as, for example, international students with L1s other than English (e.g. Chinese) or older, so-called "non-traditional" students occasionally enroll in the program. To account for this variation, KANDEL includes metadata for each participant including age, gender, major and minor subjects of study, all languages spoken and learned (including the context and duration of learning), place of birth, locations lived in, German-speaking countries traveled to, and interaction with foreign nationals.

2.2 Curricular context

The focal basic language program spans four sixteen-week-long semesters and completes the foreign language requirement for certain major subjects at this institution. Therefore, most students enroll in the program both by necessity (to complete a requirement) and by choice (by choosing German and not another foreign language out of approx. forty offered at the university). The first course in the program is designed for learners with little or no prior knowledge of German. Students with some knowledge of German (e.g., after learning it in high school) are being placed in appropriate course levels based on the results of an institutional placement test.[3]

[3] The institutional online placement test is a diagnostic test that is limited to discrete lexical and grammatical items and is not aligned with any standardized proficiency tests. In the first week of classes, instructors identify students who have been possibly misplaced and send them to the placement coordinator who conducts oral interviews with the students and corrects the placement if necessary.

The program includes multiple class sections at the same course level. Most sections follow the regular track with a uniform syllabus and textbook. However, in some semesters during the data collection period, alternative honors and business tracks were offered in addition to the regular track, which involved modifications in the syllabus. All courses in the program at the time of the corpus data collection were taught by graduate student instructors under the supervision of the author, who made all curricular decisions. The instructional approach in the program combines the communicative approach (i.e., oral interaction) with focus-on-form activities, and the program devotes an approximately equal amount of time to speaking, listening, reading, writing, vocabulary, and grammar as well as learning about the culture of German-speaking countries. Furthermore, this program has a strong learning-with-technology component including an electronic workbook and regular computer lab assignments (e.g. searching German websites for cultural information). Since all students in this program are comfortable with computers, all writing assignments are typed by students and collected electronically, including those used in the corpus collection.

The KANDEL metadata related to the curricular context include course level, cohort, semester of study, and track (regular, honors, and business).

2.3 Task variables

Short essays written by the students in response to curricular tasks rather than to external quasi-experimental tasks were collected every three to five weeks during each semester (three to five essays per semester). Some essays were written in class (typed in a computer lab) under controlled conditions (fifty minute session, without access to online or paper-based reference materials) and other essays outside of class (as homework) under uncontrolled conditions. The genres of the writing assignments were personal narratives and personal accounts (e.g. your family, your daily routine) with argumentative tasks added later on (e.g. a book review), all of which are considered level-appropriate for first and second year college-level L2 learners (Byrnes et al. 2010). Some curricular changes that were deemed necessary for the improvement of instruction were implemented over time (including a number of changes in writing assignments).[4] Although this fact reduces the direct comparability among KANDEL cohorts, it reflects the inevitable and natural variability in a real-life instructed SLA context.

The KANDEL task metadata include assignment collection date, assignment sequential number (within a cohort), writing genre, writing topic, associated textbook chapter number, location (in class / at home), and time limit.

2.4 Data collection timeline and corpus size

The data collection began in spring 2008 and ended in fall 2011, spanning five consecutive cohorts of students. “Cohort” is defined here as the class of students that entered the program in a specific semester at the novice level and progressed through four consecutive semester-long curricular levels (no summer terms were included). For example, cohort 1 includes students who progressed through four semesters of German from spring 2008 through fall 2009, and cohort 2 includes students who progressed through four semesters of German from fall 2008 through

[4] For instance, the number of required essays in the second year of study was reduced but the required length was increased. Also, to give learners an opportunity to produce more polished and sophisticated writing, more essays were assigned as homework instead of classwork.

spring 2010. The actual constituency of each cohort changed from one course level to the next (due to some students dropping out, taking a break from the program, or joining at later points via a placement test), and from one data collection point to the next (as not all students submitted all essays). Students submitted their written texts using the password-protected electronic course management system, and their ethnographic background and language learning history metadata were collected via an electronic questionnaire. Only the (anonymized) work of those students who agreed for their data to be used in research was included in the corpus. The resulting essay database contains more than 3,500 texts written by 230 participants totaling approx. 420,000 tokens (words and punctuation marks) as well as metadata for more than 20 variables. Of these raw data, all rough essay drafts from the first cohort (504 texts, 66,142 tokens) and a subset of rough essay drafts from the second cohort (185 texts, 29,635 tokens) have been annotated and entered into KANDEL, while processing of other raw data is still ongoing.[5]

The remainder of this article is devoted to the description of the data annotation process and the results of the studies conducted on the annotated corpus subsets.

3. Data annotation procedure and inter-annotator agreement

The corpus was annotated in collaboration with the corpus research team at Humboldt-Universität zu Berlin, led by Anke Lüdeling. The annotators followed the guidelines and

[5] In the data from the first two cohorts that have been annotated and entered into the corpus, no student belongs to two different cohorts (which would be the case if, for example, a student from the first cohort completed the first curricular level, then skipped a semester, and enrolled a semester later in the second curricular level with the second cohort). The corpus versions are marked with the year when the annotation was performed (e.g., v2014, v2015).

procedures developed for the L2 German corpus Falko (Lüdeling et al. 2005; Lüdeling & Hirschmann 2015; Reznicek et al. 2010; Reznicek et al. 2012). This approach uses the multi-layer standoff architecture that allows the alignment of annotations for any number of categories (see e.g. Reznicek et al. 2013 for a description). Such architecture has been used for annotating several other L2 German corpora (e.g. Gut 2012; Krummes & Ensslin 2014; Maden-Weinberger 2015; Wisniewski et al. 2013; Zinsmeister & Breckle 2012).

First, the raw learner texts were automatically tokenized, lemmatized, and annotated for parts-of-speech (POS) using the Tree Tagger tool for German (Schmid 1994) and the STTS tagset (Schiller at al. 1999). The success rate of the Tree Tagger was evaluated by comparing the tagger output to a subset of the corpus that was annotated manually. The data subset was data produced by two learners who progressed through the entire course sequence (33 texts total). This was then annotated for a subset of four tags, specifically infinitives of full verbs (VVINF), past participles of full verbs (VVPP), coordinating conjunctions (KON), and attributive adjectives (ADJA). Two human annotators independently annotated this data subset, and then resolved their disagreements (approx. 4% of the data) by discussion, to reach 100% agreement. Next, manual annotations were compared to the automatic tagger output. The high success rate of the tagger was high evidenced by the average F-score (Brants 2000) of 0.96 (see Table 1).

Table 1. Tree Tagger reliability on a KANDEL subset

POS	Tree Tagger	Annotators	Identical	Precision	Recall	F-score
VVINF	87	78	75	0.86	0.96	0.91
VVPP	74	81	74	1.00	0.91	0.95
KON	220	227	220	1.00	0.97	0.98

ADJA	125	119	119	0.95	1.00	0.98
Average				0.95	0.96	0.96

The next, most time- and effort-consuming step was manual learner error annotation using the so-called target hypothesis (henceforth, TH)[6], or corrected learner text. In LCR, it has been repeatedly pointed out that there is no one unambiguous way to annotate a learner text for TH and that, therefore, 1) different types of TH should be distinguished, 2) the criteria and procedures used in each case of TH annotation should be made explicit and transparent (Lüdeling 2008; Lüdeling & Hirschmann 2015; Reznicek et al. 2013), and 3) inter-annotator reliability should be measured and reported (Meurers 2011; Ott et al. 2012). These best practices were followed during the annotation of KANDEL as described below.

KANDEL includes multiple manual annotation layers associated with the TH. The TH1 layer is, as defined by Reznicek et al. (2013), the minimal TH that corrects only spelling and morpho-syntactic errors to form a grammatical German sentence, while the TH2 (the extended TH) layer corrects errors concerning semantics, pragmatics, and style. Furthermore, corrections performed by annotators on learner tokens (changes, deletions, insertions, and movements) are also documented in the TH1 Diff and TH2 Diff layers, for example:

(1)

Wir alles lieben zu Feier

We all love to celebration

[6] In the actual KANDEL interface, as in all corpora from the Falko family, the acronym ZH is used (that stands for the German “Zielhypothese”).

TH1	Wir	alle	lieben		Feiern		
TH1		CHA		DEL	CHA		
Diff							
TH2	Wir	alle			feiern	sehr	gerne
TH2		CHA	DEL	DEL	CHA	INS	INS
Diff							

‘We all love to celebrate.’

In this example, two learner tokens (*alles*, *Feier*) were changed (CHA) and one token (*zu*) deleted (DEL) on the TH1 level. On the TH2 level, the learner token *Feier* was changed again, and three more corrections were added: the token *lieben* was deleted and the tokens *sehr* and *gerne* were inserted (INS).

Since the primary research focus that triggered the collection and annotation of KANDEL was learner development, most available work and time resources were devoted to a reliable annotation of the longitudinal datasets collected from seventeen learners (five learners from the 1st cohort and twelve learners from the 2nd cohort) who progressed through all four consecutive semesters in the focal language program. Furthermore, since the author was primarily interested in the development of grammatical complexity in learner writing (and not, for example, discourse-level phenomena), all data currently in the corpus were annotated on the TH1 level. Only a subset of the data was annotated on the TH2 layer, as available resources permitted.

To calculate inter-annotator agreement, all texts written by five randomly selected “longitudinal” learners were annotated for TH1 and TH1 Diff by two research assistants. The two annotators were selected following Granger and Thewissen’s (2007) recommendations, with one of them being a native speaker of German (the learners' L2) and the other one a native speaker of English (the learners' L1). Both annotators had had a solid background in German linguistics and first-hand experience with the instructional context, the learners, and their L2 writing, having taught German as teaching assistants to this population for several years. The open access tool EXMARaLDA (Schmidt 2011)[7] was used that allows associating any number of annotation layers with one and the same segment of text and displaying them one under another.

To ensure the highest degree of annotation reliability possible, the following procedures were implemented. The annotators went through three rounds of independently annotating the same subsets of data for TH1 and TH1 Diff using the Falko manual guidelines (Reznicek et al. 2010) and sample annotations from the Falko corpora, whereby each round included a discussion of found discrepancies between and among the author and annotators. The inter-annotator agreement in TH1 annotations was measured using the F-score statistic (Brants 2000), which is preferable to the Kappa statistic in tasks where annotators themselves determine data spans for annotation. For example, one and the same token (or token span) may be tagged as an error by one annotator but untagged by another annotator. Therefore, we followed the approach described by Lu (2010: 486), in which "[t]wo structures are considered identical if they have the same start, end, and category label", and then the F-score is calculated based on the number of structures identified by Annotator 1 and Annotator 2 as well as identical structures. In our study, the F-score values in the three annotation rounds went up from 0.87 to 0.91 to 0.94 (see Table 2). These values, especially the final one, are considered "excellent" for data coding in SLA research (Mackey & Gass 2005: 244-245).

[7] EXMARaLDA was originally developed for annotation of oral data but its use has since been extended to written corpora.

Table 2. Inter-annotator agreement on a KANDEL subset

round	learners	texts	annotated token counts			inter-annotator agreement
			Annotator 1	Annotator 2	Identical	F-score
1	5	32	4128	4182	3600	0.87
2	5	25	3417	3420	3109	0.91
3	5	22	4081	4079	3815	0.94

During discussions, some salient disagreement patterns were discovered. After the first round, it was ascertained that many disagreements, especially regarding vocabulary, were due to over-correction: the annotators tended to correct word choice even if the learner sentence was grammatically correct. After clarifying that such corrections should be performed on the TH2 level, the number of disagreements dropped drastically. Due to the high inter-annotator agreement, the subsequent annotations were completed by one annotator, although some data subsets were then proofread by one or two annotators, especially with an eye to compliance with the new edition of the Falko annotation guidelines (Reznicek et al. 2012). Additionally, a subset of the data was also annotated on the TH2 and TH2 Diff layers. The KANDEL metadata indicate how many annotators were involved in annotating each subset, and new annotation layers can be added by other researchers as needed. Finally, all TH layers were also automatically lemmatized and POS-annotated.

4. In-depth analysis of annotator disagreements

To gain more insight into annotator disagreement, all mismatches from the inter-annotator agreement study were documented and divided into error categories (see Appendix A for examples). The question was whether the number of disagreements was higher for certain categories and data collection points. In particular, given the longitudinal nature of the data, it was interesting to see whether the annotators disagreed more about essays written at earlier or later data collection points. The following error categories were used: vocabulary, word order, noun inflection, verb inflection, spelling, and punctuation (see Appendix B for descriptive statistics). The data were normalized by 100 token counts.[8] The data were then plotted along a timeline in accordance with the sequential number of each task (i.e. data collection point) and visually inspected. Disagreements were found to occur at an overall low rate of approx. one disagreement per 100 tokens with only two distinct anomalous peaks that were qualitatively examined. One excursion in punctuation disagreements (6.7 per 100 tokens) was primarily due to a creative decision by a learner to format her essay as a dialogue separating speakers' names and turns with a semicolon. One annotator included those punctuation marks in the TH1 layer, whereas the other annotator did not. The other excursion occurred in vocabulary disagreements (7.7 per 100 tokens) due to a relatively high usage rate of English calques by learners (e.g., the

[8] A reviewer suggested using token counts for each specific category instead of the overall token counts as the normalization basis (e.g., to normalize noun inflection errors per all noun tokens). However, the main goal here is to provide an overall picture of how many annotator disagreements in different categories were found at each data collection point and to visualize this comparison by combining different categories in one graph. Therefore, a uniform normalization basis (overall token counts) is used. Ideally, one would conduct both types of analysis and use both overall counts and specific category counts as the normalization basis (see, for example, Vyatkina et al. 2015) but such a study is beyond the scope of the present research note.

German word *Klasse* (‘class’) instead of the accurate *Stunde* (‘class hour’)).[9] One annotator corrected these calques on the TH1 layer whereas the other one did not. When these two outlier values were removed to observe more general trends, the disagreement distribution became much more even across the plot.

Furthermore, it turned out that disagreement counts per error category could be divided into two groups with regard to their distribution along the timeline. These two groups are represented in separate figures for clarity. Figure 1 shows that most disagreements occurred in the vocabulary category (2.5 per 100 tokens, 1.4 – 4.1 range), fewer disagreements in the word order and noun inflection categories (approx. one per 100 tokens, 0 – 1.9 range), and the fewest disagreements in the verb inflection category (0.3 per 100 tokens, 0 – 1 range). However, disagreements in all these four categories are characterized by a uniform trend: they increase with the increase of the sequential assignment number (data collection point). This trend is illustrated more clearly in Figure 2 that plots average frequencies of disagreements in all four categories per data collection point along with a linear trendline. In contrast, no clear trend could be found for the categories of punctuation and spelling. The number of disagreements in these categories fluctuates from zero to 2.6 per 100 tokens but there is no overall increase or decrease in these frequencies depending on the sequential assignment number (Figure 3).[10]

[9] The Falko guidelines advise translating foreign words on the TH1 layer. However, it is ambiguous whether words describing foreign phenomena should be translated (e.g., ‘Fraternity’ - *Bruderschaft*).

[10] As pointed out by a reviewer, the variation in this case is too high to infer a linear trend. However, I added a linear trendline for consistency with other Figures.

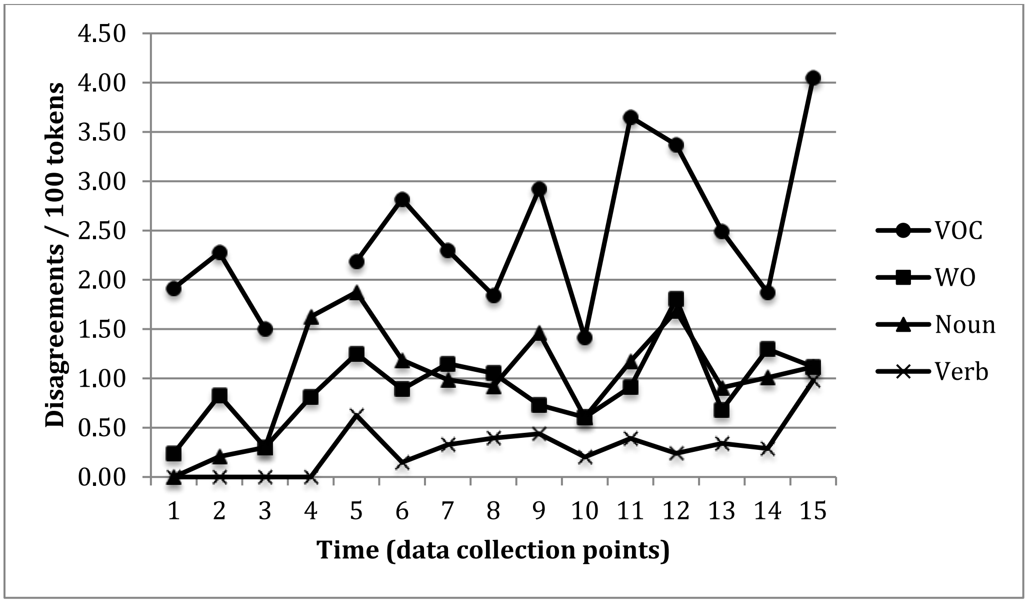

Figure 1. TH1 annotator disagreements per 100 tokens by error category: vocabulary, word order, noun inflection, and verb inflection (with the vocabulary outlier value (T4) removed)

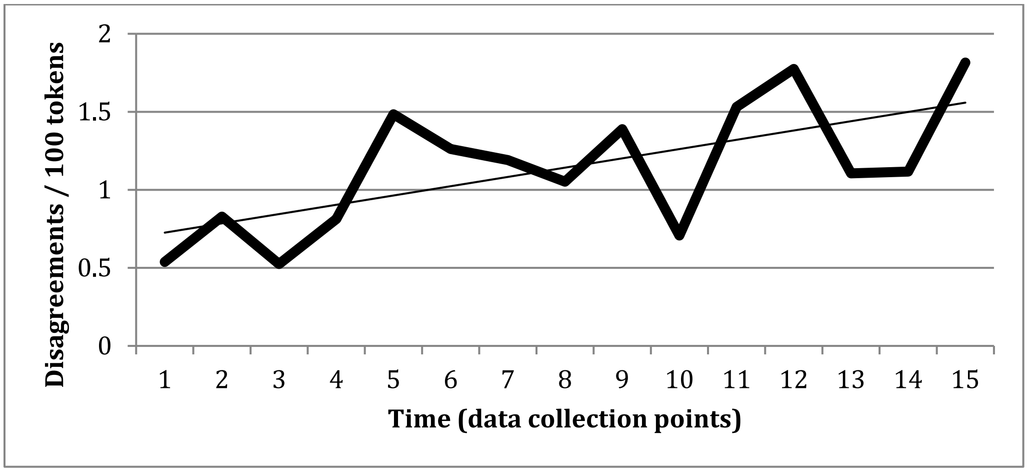

Figure 2. TH1 annotator disagreements per 100 tokens and trendline (average of vocabulary, word order, noun inflection, and verb inflection)

In addition to the TH1 layer, the error annotation categories in the TH1 Diff layer were also inspected with regard to annotator disagreements. Whereas no considerable disagreements were established in the categories “movement” and “change”, there were more disagreements in the categories “deletion” and “insertion” (see Appendix B for descriptive statistics). Moreover, there was an increasing trend for these discrepancies when plotted along the data collection timeline. As the trendlines in Figure 4 show, disagreements increased from approx. one per 100 tokens at earlier data collection points to approx. 1.5 for deletions and to approx. 3 for insertions toward the end of involved operations with the timeline. The following hypothesis may explain this finding. “Movement” and “change” only modified learner material, whereas “insertions” and “deletions” were more intrusive and, thus, involved more subjectivity on the part of the annotators. However, this hypothesis needs to be tested in an empirical study that is beyond the scope of this article.

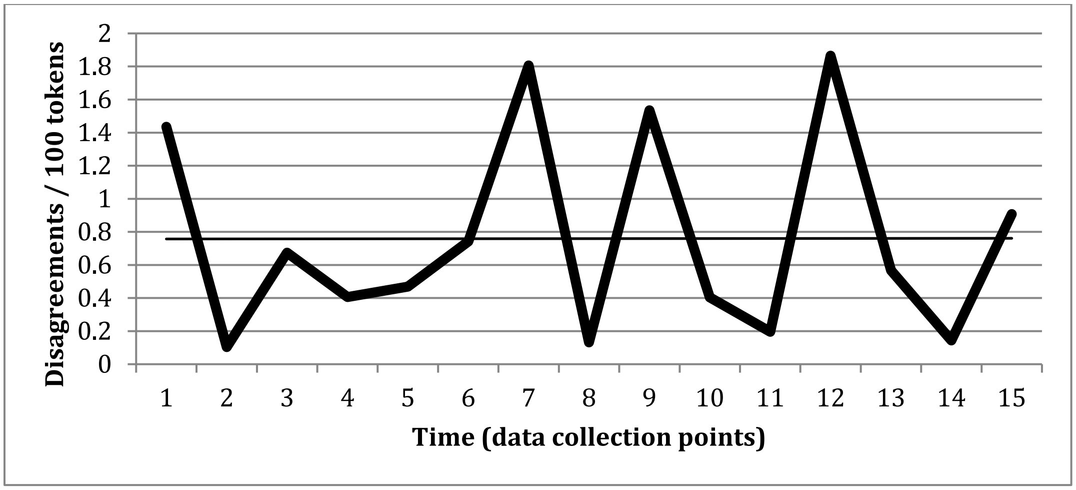

Figure 3. TH1 annotator disagreements per 100 tokens and trendline (average of spelling and punctuation)

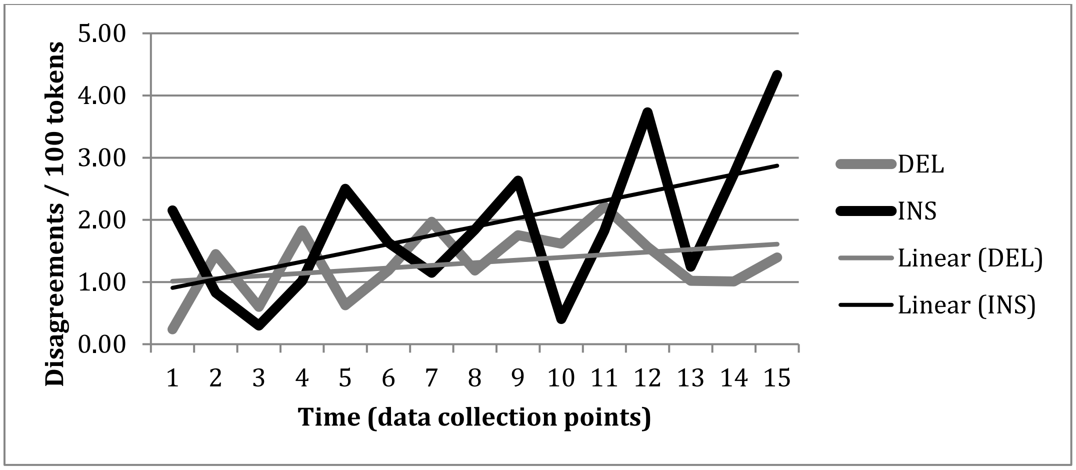

Figure 4. TH1 Diff annotator disagreements per 100 tokens

In summary, this in-depth analysis showed that some error categories are more prone to inter-annotator disagreements. The annotators in this study disagreed most on how to correct learner word choices. Another source of disagreement is inflectional morphology, a well-known stumbling block of German grammar for L2 learners. It was found that annotators agreed more on how to correct the verb inflection errors than noun inflection errors, the latter causing the same number of disagreements as German word order, another recognized area of difficulty. It must be noted that some disagreements were caused by different interpretations of the annotation guidelines and yielded divergent yet possible TH1 annotation versions (see examples 1, 2, and 6, Appendix A), whereas other disagreements were caused by annotator errors (see examples 3, 4, and 5, Appendix A).

Finally, it was found that the number of discrepancies increased while annotating essays written by learners at later time points in the curricular progression, especially when annotators deleted learner text or inserted new text. In contrast, the level of disagreement on spelling and punctuation did not depend on the data collection point. It can be hypothesized that, as learners attempt to use more complex lexical and grammatical structures over time (see Vyatkina 2012;

Vyatkina et al. 2015), tagging more complex inaccurate structures may be prone to more variety in annotators' decisions. However, this hypothesis would need to be tested in future empirical studies.

5. Completed studies and directions for future research

Several studies investigating the development of linguistic complexity in learner writing were conducted on the annotated KANDEL data. All these studies used POS annotations as surface proxies for finding and counting instances of grammatical categories (e.g. using the POS tag for subordinating conjunctions to find subordinate clauses, see Aarts & Granger 1998). Three studies used raw (i.e., not error-corrected) learner texts from the first cohort and their POS-annotations as data. Vyatkina (2012) showed that both lexical and grammatical complexity of learner writing linearly increased over time including the amount of subordination, whereas the amount of coordination linearly decreased. Vyatkina (2013a) compared frequencies of selected verb morphology POS-tags in KANDEL and in the pedagogical corpus created from the workbook used in the program. It showed that the learners gradually increased the range of different verb forms in their repertoire in general accordance with the progression found in the workbook but several divergences were also found both at the cohort and individual learner level. Vyatkina (2013b) zoomed in on two individual learners and explored developmental patterns in their use of a wider range of POS categories. The most recent study, Vyatkina et al. (2015), explored the development of syntactic modification in learner writing. It was different from the previous studies in its focus on the POS-annotations of the TH1 layer rather than of the raw learner texts.

This allowed the authors to reliably retrieve target features based on their syntactic function rather than morphological form. For example, whereas the Tree Tagger would tag attributively used adjectives missing an inflection (which is obligatory in German) as predicatively used adjectives in raw learner data, it reliably identified them on the corrected TH1 data. Working with the POS-annotated TH1 layer thus facilitated a clear focus on L2 complexity as opposed to accuracy. Future studies can combine both foci by analyzing interrelations between complexity and accuracy in learner writing. The scope can also be expanded to include semantic, pragmatic, and discourse phenomena using the TH2 annotations. Thus the multiple layers of data annotations in KANDEL allow the multidimensional and multifactorial analyses that have been strongly advocated by corpus researchers (see Gries 2015; Gries & Deshors 2015).

Furthermore, due to the fact that KANDEL comprises both longitudinal and pseudo-longitudinal data, it affords comparisons between overall group trends and developmental paths taken by individual learners. More studies in this direction could help fill a gap in LCR that tends to infer developmental trends from cross-sectional data collected from different groups of learners at different proficiency levels. Such studies are often called pseudo-longitudinal but, as Jarvis and Pavlenko (2008: 40) note in their discussion of such designs, many researchers are uncomfortable with the assumption that these studies "are capable of substituting for true longitudinal studies only to the extent that one can assume that intersubjective (and intergroup) trends across language-ability levels are similar to the trends that one would observe in a typical language user over time". KANDEL, however, presents a case that much better suits the definition of pseudo-longitudinal data. According to Jarvis and Pavlenko (2008: 40), designs similar to that of KANDEL have a higher interpretational validity than cross-sectional studies as regards developmental trends:

> For example, using an expanded multi-group design, one could collect data from a group of low beginning learners until they reached the level of low intermediate, and so forth. One could then examine and compare both longitudinal changes within groups and cross-sectional trends across proficiency levels to see how well they coincide and complement one another.

Acknowledgements

This study was supported in part by the Language Learning Small Grants Research Program, the University of Kansas General Research Fund allocation No. 2301446, and the Fulbright U.S. scholar grant, U.S.-Germany Fulbright Commission. Many thanks go to Michael Dehaven, Michael Gruenbaum, Emily Hackmann, and Melanie Piltingsrud for their help with data annotation and Anke Lüdeling, Hagen Hirschmann, Marc Reznicek, Felix Golcher, and the whole corpus research team at the Humboldt Universität zu Berlin for collaboration on this project.

Appendix A: Examples of annotator disagreements on the TH1 level by error category

Examples 1, 2, and 6 present divergent yet possible TH1 annotation versions, where disagreement was caused by different interpretations of the annotation guidelines. In contrast, examples 3, 4, and 5 include annotator errors (as marked by the asterisk).

1) Vocabulary:

Learner:	Leider werden viele Menschen **durch** den Krieg **umgeben**.
Annotator 1:	Leider werden viele Menschen **von** dem Krieg **umgeben**.

Unfortunately, many people were surrounded by the war.

Annotator 2: Leider werden viele Menschen **durch** den Krieg **umgebracht**.

Unfortunately, many people were killed by the war.

2) Word order:

Learner: **Auch ich habe** einen Stuhl.

Annotator 1: **Auch habe ich** einen Stuhl.

Annotator 2: **Ich habe auch** einen Stuhl.

I also have a chair.

3) Noun inflection:

Learner: Wilkommen zu meiner **Planet.**

Annotator 1: Wilkommen zu meinem **Planet*.**

Annotator 2: Wilkommen zu meinem **Planeten.**

Welcome to my planet.

4) Verb inflection:

Learner: ... ihre Eltern **hatte** den alten Puppenwagen flicken.

Annotator 1: ... ihre Eltern **hatte*** den alten Puppenwagen geflickt.

Annotator 2: ... ihre Eltern **hatten** den alten Puppenwagen geflickt.

…her parents repaired the old doll carriage.

5) Spelling:

Learner: meine **letze** Arbeit

Annotator 1: meine **letze*** Arbeit

Annotator 2: meine **letzte** Arbeit

my last job

6) Punctuation:

Learner: Ich habe klasse ab **9:30** bis **12:50** Uhr.

Annotator 1: Ich habe Unterricht von **9:30** bis **12:50** Uhr.

Annotator 2: Ich habe Klasse von **9.30** bis **12.50** Uhr.

I have classes from 9:30 to 12:50 o’clock.

Appendix B: Descriptive statistics for raw counts of inter-annotator disagreements per category (outlier values VOC, T4 and PUNCT, T8 removed)

Time	Tokens	VOC	WO	Noun	Verb	SP	PUNCT	DEL	INS
1	418	8	1	0	0	10	2	1	9
2	483	11	4	0	0	1	0	7	4
3	667	10	2	1	0	5	4	4	2
4	492		4	2	0	4	0	9	5
5	320	7	4	1	2	1	2	2	8
6	674	19	6	4	1	5	5	8	11
7	609	14	7	1	2	16	6	12	7
8	760	14	8	0	3	1		9	14
9	684	20	5	3	3	15	6	12	18
10	495	7	3	1	1	3	1	8	2
11	767	28	7	5	3	2	1	17	14
12	831	28	15	3	2	9	22	13	31
13	882	22	6	1	3	2	8	9	11
14	694	13	9	0	2	2	0	7	19
15	716	29	8	1	7	13	0	10	31
Average	632.80	16.43	5.93	1.53	1.93	5.93	4.07	8.53	12.40
St. Dev.	159.81	7.94	3.41	1.55	1.83	5.30	5.81	4.17	9.17

References

Aarts, J. & Granger, S. 1998. “Tag sequences in learner corpora: A key to interlanguage grammar and discourse”. In S. Granger (Ed.), *Learner English on computer*. New York: Longman, 132–141.

Alexopoulou, T., Geertzen, J., Korhonen, A. & Meurers, D. 2015. “Exploring big educational learner corpora for SLA research: Perspectives on relative clauses”, *International Journal of Learner Corpus Research* 1(1), 96–129.

Brants, T. 2000. “Inter-Annotator agreement for a German newspaper corpus”. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece: ELRA. Available at: http://www.coli.uni-saarland.de/~thorsten/publications/Brants-LREC00.pdf (accessed 4 March 2016).

Byrnes, H., Maxim, H. & Norris, J. M. 2010. “Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment [Monograph]”. *Modern Language Journal* 94(S1).

Callies, M. & Paquot, M. 2015. “An interview with Yukio Tono”, *International Journal of Learner Corpus Research* 1(1), 160–171.

Council of Europe. 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Strasbourg: Language Policy Unit. Available at: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf (accessed 4 March 2016).

Granger, S. 2015. “Contrastive interlanguage analysis: A reappraisal”, *International Journal of Learner Corpus Research* 1(1), 7–24.

Granger, S., Gilquin, G. & Meunier, F. 2015. “Introduction: learner corpus research – past, present and future”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 1–5.

Granger, S. & Thewissen, J. 2007. *Computer-aided error analysis*. Lecture presented at the Summer School *Learner Corpus Research: From corpus design to data interpretation*. University of Louvain/Belgium, 9–14 September 2007.

Gries, S. T. 2015. “Statistics for learner corpus research”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 159–181.

Gries, S. T. & Deshors, S. 2015. “EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap”, *International Journal of Learner Corpus Research* 1(1), 130–159.

Gut, U. 2012. “The LeaP corpus: A multilingual corpus of spoken learner German and learner English”. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis*. Amsterdam and Philadelphia: John Benjamins, 3–23.

Jarvis, S. & Pavlenko, A. 2008. *Crosslinguistic influence in language and cognition*. New York: Routledge.

Krummes, C. & Ensslin, A. 2014. “What’s hard in German? WHiG: a British learner corpus of German”, *Corpora* 9(2), 191–205.

Larsen-Freeman, D. 2006. “The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English”, *Applied Linguistics* 27, 590–619.

Lu, X. 2010. “Automatic analysis of syntactic complexity in second language writing”, *International Journal of Corpus Linguistics* 15(4), 474–496.

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

Lüdeling, A. & Hirschmann, H. 2015. “Error annotation systems”. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 135–157.

Lüdeling, A., Walter, M., Kroymann, E. & Adolphs, P. 2005. “Multi-level error annotation in learner corpora”, *Proceedings of Corpus Linguistics 2005*, Birmingham, UK. Available at: www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx (accessed 4 March 2016).

Mackey, A. & Gass, S. 2005. *Second language research: Methodology and design.* New York, NY: Routledge.

Maden-Weinberger, U. 2015. “‘Hätte, wäre, wenn…’: A pseudo-longitudinal study of subjunctives in the Corpus of Learner German (CLEG)”, *International Journal of Learner Corpus Research* 1(1), 25–57.

Meunier, F. & Littré, D. 2013. “Tracking learners’ progress: adopting a dual corpus cum experimental data approach”, *Modern Language Journal* 97(S1), 61–76.

Meurers, D. 2011. On automatically analyzing learner language. Keynote lecture presented at *Learner Corpus Research* 2011, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 15-17 September 2011. Available at: http://www.sfs.uni-tuebingen.de/~dm/handouts/louvain-11-09-17.pdf (accessed 4 March 2016).

Ortega, L. & Byrnes, H. 2008. “Theorizing advancedness, setting up the longitudinal research agenda”. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* New York, NY: Routledge/Taylor & Francis, 281–300.

Ortega, L. & Sinicrope, C. 2008. *Novice proficiency in a foreign language: A study of task-based performance profiling on the STAMP test*. (Technical report). University of Oregon, Center for Applied Second Language Studies.

Ott, N., Ziai, R. & Meurers, D. 2012. “Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context”. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis*. Amsterdam and Philadelphia: John Benjamins, 47–69.

Reznicek, M., Lüdeling, A. & Hirschmann, H. 2013. “Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture”. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data*. Amsterdam and Philadelphia: John Benjamins, 101–124.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H. & Andreas, T. 2012. *Das Falko-Handbuch: Korpusaufbau und Annotationen, Version 2.01*. Available at: https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01 (accessed 4 March 2016).

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C. & Andreas, T. 2010. *Das Falko-Handbuch: Korpusaufbau und Annotationen, Version 1.0.1*. Available at: https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v1.0.1 (accessed 4 March 2016).

Schiller, A., Teufel, S., Stöckert, C. & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS* [Guidelines for tagging German corpora of written language with STTS]. Technical Report. Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung [Institute for Machine Language Processing].

Schmid, H. 1994. "Probabilistic part-of-speech tagging using decision trees", *Proceedings of the international conference on new methods in language processing*. Manchester, UK, 44–49. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1139&rep=rep1&type=pdf (accessed 4 March 2016).

Schmidt, T. 2011. "A TEI-based approach to standardising spoken language transcription", *Journal of the Text Encoding Initiative* 1. Available at: http://jtei.revues.org/142 (accessed 4 March 2016).

Vyatkina, N. 2012. "The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study", *Modern Language Journal* 96(4), 576–598.

Vyatkina, N. 2013a. "Analyzing part-of-speech variability in a longitudinal learner corpus and a pedagogic corpus". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead. Corpora and Language in Use - Proceedings 1*. Louvain-la-Neuve: Presses universitaires de Louvain, 479–491.

Vyatkina, N. 2013b. "Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus", *Modern Language Journal* 97(s1), 11–30.

Vyatkina, N. 2016. "Data-driven learning for beginners: The case of German verb-preposition collocations", *ReCALL* 28(2), 207-226. doi: 10.1017/S0958344015000269

Vyatkina, N., Hirschmann, H. & Golcher, F. 2015. "Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study", *Journal of Second Language Writing* 29, 28–50.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., Abel, A. & Hana, J. 2013. "MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data". In: *ICT for Language Learning, Conference Proceedings 2013*. Libreriauniversitaria.it Edizioni. Available at: http://conference.pixel-online.net/ICT4LL2013/common/download/Paper_pdf/322-CEF03-FP-Wisniewski-ICT2013.pdf (accessed 4 March 2016).

Zinsmeister, H. & Breckle, M. 2012. "The ALeSKo learner corpus: Design – annotation – quantitative analyses". In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis*. Amsterdam and Philadelphia: John Benjamins, 71–96.