KoKo: an L1 Learner Corpus for German

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, Egon Stemle
Institute for Specialised Communication and Multilingualism
European Academy of Bozen/Bolzano
andrea.abel@eurac.edu, aivars.glaznieks@eurac.edu,
lionel.nicolas@eurac.edu, egon.stemle@eurac.edu

Abstract

We introduce the KoKo corpus, a collection of German L1 learner texts annotated with learner errors, along with the methods and tools used in its construction and evaluation. The corpus contains both texts and corresponding survey information from 1,319 pupils and amounts to around 716,000 tokens. The evaluation of the performed transcriptions and annotations shows an accuracy of orthographic error annotations of approximately 80% as well as high accuracies of transcriptions ($> 99\%$), automatic tokenisation ($> 99\%$), sentence splitting ($> 96\%$) and POS-tagging ($> 94\%$). The KoKo corpus will be published at the end of 2014. It will be the first accessible linguistically annotated German L1 learner corpus and a valuable source for research on L1 learner language as well as for teachers of German as L1, in particular with regards to writing skills.

Keywords: L1 Learner Corpus, Writing Skills, Corpus Evaluation

1. Introduction

Using linguistically annotated corpora in learner language research has received growing attention over the past 20 years (Granger et al., 2013). The field of learner corpus linguistics usually defines learner corpora as "systematic computerized collections of texts produced by language learners" (Nesselhauf, 2005). Learner corpora are usually annotated with the help of a standardized system of error tags (Díaz-Negrillo and Domínguez, 2006). Learner corpora also tend to provide meta-information, such as the authors' L1, age, gender, etc. and other valuable information from all relevant levels of linguistic description. Usually, linguistic annotations are performed automatically when automatic processing reaches a certain accuracy, which is often the case for lemma and part-of-speech (POS) information. From a technical perspective, the annotations are either inline (Granger, 2003) or in a multi-layered way using a stand-off format (Lüdeling et al., 2005; Reznicek et al., 2013; Zinsmeister and Breckle, 2012; Hana et al., 2010; Hana et al., 2012).

In this paper, we refer to people as L1 learners when they are still in the process of learning their L1 or related skills of importance such as writing and text production. We also refer to people as L2 or Foreign Language (FL) learners when the language concerned is not their L1. Prototypically, instances of L1 learner language can be found in the educational and academic context. From a linguistic point of view, the texts written by L1 language learners are likely to have many features of non-standard writing in common with L2/FL learners. However, since some features are specific to either L1 or L2/FL learners, both learner types relate to separate learner varieties. From the perspective of computational processing, L1 and L2/FL learner corpora are fully equivalent since both are compilations of textual data that may deviate from the standard variety.

Analysing linguistically annotated texts produced by language learners offers insight into both their competences and their difficulties, which in turn can be used to improve their competences accordingly. Teachers and researchers can also rely on learner corpora to observe key obstacles in language development and derive, with reliable grounding, guidelines for didactics in language teaching (Aijmer, 2009). Furthermore, learner corpora represent useful resources in the context of language assessment and certification, e.g. for illustrating the reference levels and descriptors of the Common European Framework of Reference for Languages (Council of Europe, 2001; Abel et al., in print; Hawkins and Filipović, 2012; Spinelli and Parizzi, 2010). Compiling learner corpora is therefore an approach adopted by numerous past and ongoing research projects in the fields of language acquisition and teaching, especially projects focusing on foreign language.

Learner corpora also provide valuable data for the development and evaluation of natural language processing tools for learner language (Meurers, 2013). Their annotations can be used as input data for automatic tasks such as proficiency classification (Hancke et al., 2012; Vajjala and Meurers, 2012) and, when dealing with L2 learners, native language identification (Jarvis et al., 2012) whereas the textual data can be used as grounding data for developing tools and applications assisting language learners in improving their skills.

In this paper, we report on ongoing research that is concerned with the creation of an L1 learner corpus for German. The corpus currently consists of refined transcriptions that have first been annotated with iteratively refined surface annotations, i.e. annotations of the surface form with limited linguistic interpretation, and later extended with higher-level annotations requiring linguistic context and interpretation.

The paper is structured as follows. In section 2., we provide an overview of related work on learner corpora. This is followed by a presentation of the KoKo corpus, including details on its origins, methodology of data collection, annotation, size, and contents (section 3.). In section 4., we present the results on an evaluation of the corpus. Finally, section 5. addresses some aspects of future work.

2. Related Work

Annotated L1 learner corpora that were build with the purpose of analysing German L1 learner language are still rare.[1] The related work can be roughly divided into two groups: research that includes German L1 reference corpora and research concerned with pupils' L1 writings.

2.1. German L1 Reference Corpora

Some linguistically annotated collections of L1 writings serve as reference corpora in German L2/FL corpora. In the Falko corpus (Reznicek et al., 2012), for example, there are 57 summaries (around 21,000 tokens) and 95 essays (around 70,000 tokens) written by German native speakers. The majority of the essays (around 59,000 tokens) were written by pupils attending final grades at secondary schools (between 17-19 years old) while the rest was written by university students. The essays were annotated for orthographic, grammatical and lexical errors. All L1 texts are annotated for POS and lemmas.

A reference corpus of German L1 writers for argumentative essays (around 12,000 tokens) also exists in the Kobalt corpus, a collection of German FL learner texts written by learners with either Russian, Swedish, or Chinese as L1 (Zinsmeister et al., 2012). Both the Falko and the Kobalt corpus can be queried via the Falko interface ANNIS (Zeldes et al., 2009), an enhanced infrastructure for corpus queries and analyses.[2]

The limits for studying L1 learner language in such corpora are manifold. First, they were not built to analyse L1 language competences but to contrast them with the writings of L2/FL learners. Therefore, annotations focus mainly on the performance of L2/FL learners as captured in specific errors (e.g. errors in gender assignment) that are rather rare in L1 learners' language. Second, the size of such corpora is usually small. Hence, they hardly serve for investigations on L1 learner language in a broad sense as findings can rarely be generalized.

2.2. German Corpora of Pupils' L1 Writings

2.2.1. Accessible Corpora

For younger pupils with German as L1, we are aware of one partly annotated text corpus built as a test corpus for a tool that automatically annotates orthographic errors (Thelen, 2010).[3] The corpus was created from around 750 narrations of a picture story produced by second graders. Due to methodological shortcomings of the study, the corpus is hardly usable for research that goes beyond research on orthographic and grammatical errors.

Some text collections are provided to the public in data formats such as pdf or rtf. For example, Augst et al. (2007) provide collections of 5 different text types (narration, instruction, report, description, argumentation) gathered during a longitudinal study on the development of writing skills.[4] 39 pupils have produced such texts in three successive grades (2nd, 3rd, and 4th grade), 16 pupils of them participated also in the sixth grade. At each step of the data collection, participants followed the same instructions. The aim of the data collection was to describe the pupils' advancements in writing over time.

Another corpus in raw data format is the so-called "Ludwigsburger Aufsatzkorpus", which is provided on CD by Fix and Melenk (2002). It consists of around 2300 texts written by approximately 650 sixth graders. There are two types of texts related to each writer: summaries and texts that were induced by an image. Although accessible, no infrastructure is provided for corpus queries and analyses, and neither the "Ludwigsburger Aufsatzkorpus" nor Augst et al.'s collection is enriched with linguistic annotations.

2.2.2. Non-accessible Corpora

Most collections of German L1 learner texts are not accessible. Among them is the "Heidelberger Korpus Schülertexte (HEIKOS)" (Berg et al., 2010). It consists of written narrations of a picture story collected from sixth-graders of diverse German secondary schools or schools for children with speech and language deficits. All narrations were manually annotated for grammatical errors focusing on errors regarding nominal and verbal inflection. The purpose of the study was to analyse the linguistic heterogeneity of sixth-graders in the realms of writing, reading, narrating, and in conversation.

Another collection of German L1 learner texts, which is not publicly accessible, is introduced in Klieme (2008). The collection comprises letters of complaint and private letters written by around 10,000 pupils in 9th grade. All letters were rated for levels of competences. Raters had to consider both formal correctness of the language (e.g. orthographic and syntactic errors) as well as semantic and pragmatic appropriateness. The letters were part of a larger study to estimate the conditions in the German and English classroom by analysing the productive and perceptive linguistic skills in German and English of pupils in Germany (DESI).

Finally, Hanser et al. (1994) report on a corpus which was collected to analyse language skills of secondary-school graduates in German-speaking Switzerland. The authors used an analysis grid (Nussbaumer and Sieber, 1994) to annotate errors, functional and stylistic appropriateness as well as contents. Although the collection is referred to as digitized, it is not accessible to the research community.

3. The KoKo Corpus

3.1. Origins & Objectives

The Koko corpus is a key outcome of the Koko project that in turn is part of Korpus Südtirol (Abel and Anstein, 2011). The latter is a corpus linguistic initiative to collect, file and process texts in order to make them available to the public and document the use of written German in South Tyrol.

The KoKo corpus has been created with the aim to investigate and describe the writing skills of German-speaking secondary-school pupils at the end of their school career by analysing authentic texts produced in classrooms.

[1] For American English see for example O'Donnell and Römer (2012) and Römer and O'Donnell (2011).

[2] https://korpling.german.hu-berlin.de/falko-suche/

[3] https://repositorium.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-201006096307

[4] https://www.text-sorten-kompetenz.de/

The corpus building process was guided by two goals:

1. to describe writing skills at the transition from secondary school to university,
2. to determine external factors that may influence the distribution of writing skills, such as the region, sociolinguistic (gender, age), socio-economic, and language-related biographical factors (L1, preferred variety of German, reading and writing habits, etc.).

3.2. Data Collection

In May 2011, 1,511 pupils from 85 classes and 66 schools participated in the Koko project by writing a text and answering a questionnaire gathering background information.

3.2.1. Participants

The pupils were selected from three different German-speaking areas: North Tyrol (Austria), South Tyrol (Italy), and Thuringia (Germany). At the time of data collection, all writers attended secondary schools one year before their school-leaving examinations. Classes were sampled randomly, using the size of the cities in which the schools were located (small vs. medium vs. big) and the type of school (providing general education vs. education specific to a particular profession) as strata for the sampling. Since data were collected during regular courses, the typical formation of secondary-school classes in the three regions is represented in the whole corpus. Most of the participants are German native speakers (n=1319, 82.7%). In addition, pupils with the following L1s have participated: Italian (n=28, 1.8%), Ladin. (n=19, 1.2%), Russian (n=9, 0.6%), English, Turkish (n=8, 0.5%), Serbian (n=5, 0.3%), Albanian (n=4, 0.3%), Bosnian, Chinese, Vietnamese (n=3, 0.2%), Dutch, French, Spanish (n=2, 0.1%), Armenian, Bulgarian, Croatian, Czech, Finish, Greek, Hindi, Japanese, Korean, Portuguese, Tagalog, Ukrainian, Urdu, Slovakian (n=1, 0.1%). 110 pupils (7.3%) did not disclose their L1.

3.2.2. Survey Areas

There are several reasons for having selected pupils from North Tyrol, South Tyrol and Thuringia.

From a linguistic perspective, pupils from these regions are suitable to be contrasted with each other. The three regions are part of the German-speaking language area; beyond the differences of the respective standard variety in each region (Clyne, 1992; Ammon et al., 2004), the regions differ in further noticeable aspects: whereas South Tyrol is a plurilingual region (German, Italian, Ladin)[5] with a vivid use of both the South Tyrolean dialect and the German standard, North Tyrol is a predominant monolingual area in which the dialect and the standard variety are widely used. Thuringia by contrast is a predominant monolingual region where the old dialects are not used any more, particularly by young speakers who rather prefer a regional vernacular that is close to the standard language. The differences regarding the linguistic situation in the three regions are suitable for controlling any possible influence of a plurilingual environment and of being a dialect speaker on pupils' writing skills in the standard variety.

From a demographic perspective, the three regions are comparable. In particular, North Tyrol and South Tyrol are very similar. In both regions, there are no big cities (the largest ones have around 100,000 inhabitants), and the total number of citizens is comparable (appr. 650,000 vs. 500,000 respectively). Compared to the other two regions, Thuringia has more inhabitants: approximately 2.2 million people. Although there are three bigger cities with more than 100,000 inhabitants, half of the citizens live in settlements with a population under 10,000, which is comparable to the situation in South Tyrol (56%) and North Tyrol (66%).

3.2.3. Method

As part of the regular course work, pupils were asked to write an argumentative essay in class on one and the same topic. All essays were originally handwritten, and transcribed later.

The reason for collecting handwritten texts was that all texts should be produced with the same means. Using a pen or a keyboard influences the text production which makes it impossible to compare handwritten text with those written on a computer with respect to the authors' writing skills (Grabowski et al., 2007). As computers were not available in most classes, all texts were produced by hand.

In addition to the text production task, all participants completed a written survey with sociolinguistic, socio-economic and language-related biographic questions. From 1,511 pupils, 1,503 essays were manually transcribed and the data of the corresponding written surveys transferred into a spreadsheet. The subset of texts written by German native speakers constitute the KoKo L1 corpus. It contains texts and corresponding survey information from 1,319 pupils and is roughly equally distributed over the three regions, amounting to a total of 716,405 tokens (version KoKo2, Dec 2012).

sub-corpus	all learners		L1 only	
region	tokens*	texts	tokens*	texts
North Tyrol	233,098	457	206,439	404
South Tyrol	222,209	520	192,891	451
Thuringia	353,674	521	317,075	464
unknown	2,349	5	-	-
total	811,330	1503	716,405	1,319

* without punctuation

Table 1: Sub-corpora within the KoKo corpus (version KoKo2, Dec 2012) according to region and L1.

The average word length in the corpus is 5.38 letters, the average sentence length is 17.39 words, and the average text length is 543.14 words.

[5] In South Tyrol, German is legally equal to Italian which implies that children with L1 German have the right to be educated in German from nursery to secondary school level. For a better understanding of the linguistic situation in the educational system of South Tyrol see Abel et al. (2012).

3.3. Format and Tools

The hand-written learner essays were scanned and then transcribed using XMLmind[6] with a custom style sheet allowing us to perform annotations on-the-fly (section 3.4.2.). Since the remaining set of annotations required the ability for the annotations to overlap and to span non contiguous sequences of tokens, we later converted the corpus to a stand-off format. We studied several freely available multipurpose formats associated with a dedicated editor and decided for MMAX2 (Müller and Strube, 2006), a stable tool used in many past and ongoing projects.

In order to be able to recover any manipulation error, we decided to rely on the well supported revision control system subversion[7]. This also has the advantage of making the data exchange simpler and less error-prone.

In order to facilitate linguistic processing of the data, all annotations had to be accessible through an integrated user-friendly interface with the ability to formulate sophisticated queries. ANNIS, an open source search and visualisation architecture for complex multi-level linguistic corpora, fulfilled our demands.

For converting the corpus from the MMAX2 format to the ANNIS format, we relied on the conversion suite SaltNPepper (Zipser and Romary, 2010) for which a plug-in for both MMAX2 and ANNIS format exists.[8]

3.4. Annotation Schema

The annotations of the Koko corpus can be grouped into three types: metadata of texts, manual annotations, and automatic annotations.

3.4.1. Metadata

Metadata was manually extracted and transcribed from the questionnaires. The metadata currently consists of five types of information (version KoKo2, Dec 2012): the writers' L1, the type of school, the region of origin, the writers' gender, and the grade attended at data collection.

The main objective for collecting this information is to perform sociolinguistic analyses through the detection of relations between non-linguistic information and text features such as text length, sentence length, lexical variation, etc. In addition, the corpus can also be used for more sophisticated statistical analyses relating sociologically relevant information with features obtained from the linguistic annotations. Considering the linguistic situations (cf. section 3.2.2.), for example, we would like to find out if the region of origin (South Tyrol vs. North Tyrol vs. Thuringia) has an impact on the pupils' writing skills and, if yes, in which way. Other explanations for the distribution of writing skills (e.g. type of school and socio-economic background of the family) have to be considered as well.

3.4.2. Manual Annotations

Manual annotations were performed in several passes on the basis of a specifically crafted tag set and a detailed annotation manual.

Transcription Annotations. During the transcription of the handwritten documents, the corpus was manually annotated with surface features of the text, such as graphical arrangement (outline and other pretext elements, title, paragraphs, emphasis, footnotes, and postscript elements) and self-corrections (insertions, deletions). Emoticons and symbols were also annotated. Finally, the transcriber had to annotate those words that could not be read (unreadable) or cleary identified (ambiguous). The annotator could also comment on words or parts of the text whenever s/he thought it could contribute to a better understanding of the transcription (comment). An overview of the numbers of annotations is provided in table 2.

annotation	numbers	mean per text
deletion	15,313	11.61
paragraph	11,218	8.50
insertion	5,354	4.06
unreadable	2,944	2.23
title	658	0.50
emphasis	389	0.29
comment	233	0.18
symbol	140	0.11
alternative	76	0.06
ambiguous	38	0.03
pretext	21	0.01
emoticon	14	0.01
outline	10	0.01
postscript	8	0.01
footnote	2	0.00

Table 2: Annotations performed during the transcription phase (version KoKo2, Dec. 2012).

Linguistic Annotations. Two dimensions of annotations of errors as well as other linguistic features have been considered: (a) linguistic category such as orthography or grammar allowing also for subcategories (e.g. word order regarding grammar) and (b) target modification classification (e.g. omission, addition) (Díaz-Negrillo and Domínguez, 2006). Furthermore, error correction (formulation of a target hypothesis) has been inserted as a further dimension of the manual annotation (Lüdeling et al., 2005). The following linguistic dimensions are included in the annotation scheme: orthography, grammar, lexis and several aspects at the textual level. For example, on the orthographic level specific deviations (orthographical errors, punctuation errors) from the standard written variety of German were carefully annotated, classified, and given a target hypothesis on a separate level. The classification schema for orthographical errors comprises 28 distinct categories that can be assorted to seven superordinate categories. The categories are based on the rules and principles of German orthography (Duden, 2005; Duden, 2006; Fuhrhop, 2005):

1. upper and lower case errors,
2. separate and compound spelling errors,
3. omission of letters,

[6] http://www.xmlmind.com/xmleditor/

[7] http://subversioning.tigris.org/

[8] For a detailed description of the workflow see Glaznieks et al. (2014).

4. adding of letters,

5. confusion of letters,

6. special cases: missing or false use of apostrophes, errors within abbreviations, misspelled proper names.

An overview of the numbers of annotations in the German L1 sub-corpus is provided in table 3.

superordinate category	numbers	mean per text
1. upper/lower case errors	3,970	3.01
2. separate/compound spelling errors	2,497	1.89
3. omission of letters	2,134	1.62
4. adding of letters	882	0.67
5. confusion of letters	1,166	0.88
6. special cases	934	0.71

Table 3: Numbers of orthographic errors by superordinate categories (L1 corpus, version KoKo2, Dec. 2012).

A sub-sample of the KoKo corpus (597 texts selected according to the above mentioned strata (see section 3.2.1.) and equally distributed over the three regions) was additionally annotated for grammatical errors. The annotation schema for grammatical errors is based on the characteristics of German grammar (Duden, 2005; Zifonun et al., 1997). It is comparable to the one used in the Falko corpus as it covers most of the grammatical phenomena that are described in Reznicek et al. (2012). Differences are motivated by the type of learner (L1 vs. L2). The annotation schema consists of the following annotations.

(a) Correspondence Relations

  (a1) *Correspondence*: erroneous selection of case, number, gender or person of a dependent word with respect to government and congruency (always in combination with the annotation correspondence referent).

  (a2) *Correspondence referent*: the governing word or head of a phrase in case of an erroneous selection of case, number, gender or person with respect to government and congruency (always in combination with the annotation correspondence).

(b) *Inflection*: incorrect inflected forms that are independent of a governing element, e.g. forms following the wrong inflection paradigm such as weak instead of strong verbal inflection.

(c) *Incompleteness*: incomplete sentences and phrases as well as the incorrect use of ellipses.

(d) *Redundancy*: erroneous repetitions of words and parts of sentences.

(e) *Anacoluthon*: ungrammatical blending of phrases and clauses.

(f) *Word order*: violations of any kind of word order restrictions.

(g) *Not categorisable grammatical error*.

Most grammatical annotations (a-g) have different subclassification schemas, with varying numbers of subcategories. For example, annotation of (e) contains the subcategories of: (e.1) syntactic errors that occur due to a transposition of the initially intended syntactic structure within one clause, and (e.2) syntactic errors that occur due to a retraction of the initially intended syntactic structure by self-correction. On the contrary, the annotation of (a1) requires a specification of the kind of the error that in turn depends on the part of speech of the item to be annotated. As a consequence, 13 categories were created, each with a distinguishing subcategorisation that specifies the error; for example, the category (a1.1) adjective is divided in five subcategories specifying the following correspondence errors: (a1.1a) false case, (a1.1b) false number, (a1.1c) false gender, (a1.1d) false inflection paradigm, and (a1.1e) unknown (if it is not determinable).

The grammatical annotations will be part of the next version of the corpus (version KoKo3, end of 2014). Table 4 shows preliminary numbers of the grammatical annotations on the sub-sample of 597 texts.

annotation		numbers	mean per text
(a1)	Correspondence	1,697	2.84
(a2)	Correspondence referent	1,572	2.63
(b)	Inflection	246	0.41
(c)	Incompleteness	381	0.64
(d)	Redundancy	69	0.11
(e)	Anacoluthon	127	0.21
(f)	Word order	110	0.18
(g)	Not categorisable gram. error	111	0.18

Table 4: Preliminary numbers of grammatical error annotations in a sub-sample of 597 texts.

3.4.3. Automatic Annotations

Automatic annotations, namely, tokenisation, sentence splitting, POS-tagging and lemmatisation were done with the help of the IMS TreeTagger (Schmid, 1994). To achieve a higher precision, we used the target layer of the corpus to avoid tagging errors due to misspelled words.

Indeed, the performances of the vast majority of natural language processing tools depend on how similar the texts are to the language they have been trained on or devised for. Therefore, best performances are usually achieved with texts written in the standard variety of the language. Thus, performances tend to drop when texts deviate from the standard variety. By exchanging any error-annotated tokens with the corresponding target form, we generated a version of the KoKo corpus that, with respect to orthography, barely deviated from standard German, and thus achieved an accuracy that is within the range of state-of-the-art POS-tagging performance for German (cf. section 4.2.).

level	total size		correct		accuracy in %	
	token	sentence	token	sentence	token	sentence
(1) transcription	4,842	255	4,825	238	99.65	93,33
(2) orthographic errors	61	49	49	40	80.33	86.96
(3) tokenisation	4,842	255	4,841	254	99.98	99.60
(4) sentence splitting	-	255	-	247	-	96.86
(5) POS-tagging	4,191	227	3,969	96	94.70	42.29

Table 5: Evaluation of the quality of the KoKo corpus (version KoKo2, Dec 2012).

4. Corpus Evaluation

4.1. Evaluation Procedure

Inspired by the Agile Corpus Creation approach (cf. Voormann and Gut 2008) we iteratively performed quality checks for the manual annotations during the transcription and the annotation phases. These phases dealt with transcription and annotation errors that affect the automatic annotation of the data. To identify errors we focused on out-of-vocabulary words from the IMS TreeTagger.

In order to evaluate the final corpus, a random sample of 255 sentences (about 4,800 tokens) representing 0.54% of the 46,734 sentences of the error annotated corpus was evaluated in order to create a gold standard. The evaluation was done for (1) the transcription, (2) the orthographic error annotations, (3) the tokenisation, (4) the sentence splitting, and (5) the POS-tagging. We calculated these accuracies because they influence the usability of the corpus for linguistic research and they are indicative of its overall quality. With respect to (1), all 255 sentences of the sample were compared with the scanned handwritten original version of the texts and transcription errors were counted. For the evaluation of (2), all sentences were checked again for orthographic errors. Annotated orthographic errors were verified and missing annotations were marked. The performance of (3, 4) was evaluated on the basis of the output of the IMS TreeTagger. The output was manually checked and errors were counted. On the basis of the evaluation of (1-4) 28 sentences with transcription errors (17 sentences), flaws in manual annotations on the orthographic level (9), errors in tokenisation (1) and sentence splitting (8) were excluded from the subsequent evaluation.[9] The remaining 227 sentences were then evaluated with respect to (5), the POS-tagging errors: the POS-tagging output was manually checked by two independent annotators (disagreements were discussed until a conclusion was reached), and the corrected POS tags were added on a separate layer, now constituting our gold standard.

4.2. Evaluation Results

Accuracy in the dimensions (1-5) varies (cf. table 5): with an accuracy rate of 99.6% the transcription (1) on the sample is very accurate. The accuracy rate of orthographic error annotations (2), by contrast, is lower as expected; it reached an accuracy rate of around 80%. However, we are not aware of any numbers for comparison that could help to evaluate this result. With respect to the automatic processing of the sample, one tokenisation error (3) remained in the sample, which leads to an accuracy rate of 99.9%. Sentence splitting (4) worked quite accurately; although they should have been split, three sentences were not, and five were wrongly split. The POS-tagging accuracy of 94.7% (3) is on the lower end of the state-of-the-art POS-tagging performance for German, which reaches up to 97% (Schmid, 1995). Taking into account that the target layer does not include grammatical errors, the token level accuracy is excellent. However, it should be noted that only 42.3% of the sentences are free of tagging errors. The low accuracy rate on sentence level is most likely due to some well-documented shortcomings of the TreeTagger (Schmid, 1995) that usually appear once per sentence, which probably causes the errors to be equally spread over the sample. So far, automatic adjustments of such errors have not been possible.

[9] In seven sentences two different types of errors occurred.

5. Future Work and Conclusion

We intend to make the corpus accessible via ANNIS by the end of 2014. In order to make the corpus available to a larger public, we are now considering the possibility of releasing version 3 of the corpus in Paula format, a stand-off format that has similar properties to the MMAX2 format but has originally been designed to be an exchange format for linguistic content. As such, it is able to represent a wider range of annotations more efficiently.

In the future, further meta data regarding language biography and language use will be added. Currently, the sub-sample of 597 texts (cf. section 3.4.2.) that has been annotated for grammatical errors is also being evaluated according to several aspects of text quality (e.g. cohesion, coherence, internal structure and composition of the text) using an evaluation sheet. They will also be annotated for phenomena on the lexical level (e.g. semantic errors, incorrect use of formulaic sequences).

This paper introduced the KoKo corpus, a collection of German L1 learner texts annotated with learner errors, along with the methods and tools used in its construction.

Since comparable collections of texts written by pupils are either not accessible, have not been enriched with linguistic information, or, although accessible, are only partly annotated, the KoKo corpus will be the first accessible linguistically annotated German L1 learner corpus. The corpus will be a valuable source for research on L1 learner language, in particular for the research on writing skills, and for teachers of German as L1, in particular for the teaching of L1 German writing skills.

6. References

Abel, A. and Anstein, S. (2011). Korpus Südtirol-Varietätenlinguistische Untersuchungen. In Abel, A. and Zanin, R., editors, *Korpusinstrumente in Lehre und Forschung*, pages 29–54. University Press Bozen.

Abel, A., Vettori, C., and Forer, D. (2012). Learning the neighbours language: the many challenges in achieving a real multilingual society. In European Centre for Minority Issues and European Academy Bozen/Bolzano, editor, *European Yearbook of Minority Issues*, pages 271–304. Martinus Nijhoff Publishers.

Abel, A., Wisniewski, K., Nicolas, L., Boyd, A., Hana, J., and Meurers, D. (in print). A trilingual learner corpus illustrating European reference levels. In *Proceedings of the Learner Corpus Research 2013 Conference (LCR2013)*.

Aijmer, K., editor. (2009). *Corpora and Language Teaching*. John Benjamins.

Ammon, U., Bickel, H., Ebner, J., Esterhammer, R., Gasser, M., Hofer, L., Kellermeier-Rehbein, B., Löffler, H., Mangott, D., Moser, H., Schläpfer, R., Schlossmacher, M., Schmidlin, R., and Vallaster, G., editors. (2004). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. de Gruyter.

Augst, G., Disselhoff, K., Henrich, A., Pohl, T., and Völzing, P. (2007). *Text-Sorten-Kompetenz*. Peter Lang.

Berg, M., Berkemeier, A., Funke, R., Glück, C., Hofbauer, C., and Schneider, J. (2010). Sprachliche Heterogenität in der Sprachheil- und der Regelschule.

Clyne, M. G. (1992). German as a pluricentric language. In Clyne, M. G., editor, *Pluricentric languages. Differing norms in different nations*, pages 117–148. De Gruyter.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Council of Europe Publ.

Díaz-Negrillo, A. and Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, 19:83–102.

Duden. (2005). *Die Grammatik*. Bibliographisches Institut.

Duden. (2006). *Die deutsche Rechtschreibung*. Bibliographisches Institut.

Fix, M. and Melenk, H. (2002). *Schreiben zu Texten-Schreiben zu Bildimpulsen*. Schneider-Verlag Hohengehren.

Fuhrhop, N. (2005). *Orthografie*. Universitätsverlag Winter.

Glaznieks, A., Nicolas, L., Stemle, E., Abel, A., and Lyding, V. (2014). Establishing a standardised procedure for building learner corpora. *Apples Journal of Applied Language Studies*, 8(1).

Grabowski, J., Blabusch, C., and Lorenz, T. (2007). Welche Schreibkompetenz? Handschrift und Tastatur in der Hauptschule. In Becker-Mrotzek, M. and Schindler, K., editors, *Texte schreiben*, pages 41–62. Gilles und Francke.

Granger, S., Meunier, F., and Gilquin, G. (2013). *Twenty Years of Learner Research: Looking Back, Moving Ahead*, volume 1. Presses universitaires de Louvain.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO journal*, 20(3):465–480.

Hana, J., Rosen, A., Škodová, S., and Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 11–19. Association for Computational Linguistics.

Hana, J., Rosen, A., Stindlová, B., and Jäger, P. (2012). Building a learner corpus. In *LREC12*, pages 3228–3232.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for german using lexical, syntactic, and morphological features. In *COLING*, pages 1063–1080.

Hanser, C., Nussbaumer, M., and Sieber, P. (1994). Was sich in geschriebenen Texten zeigt. In Sieber, P., editor, *Sprachfähigkeiten–besser als ihr Ruf und nötiger denn je*, pages 187–301. Sauerländer.

Hawkins, J. A. and Filipović, L. (2012). *Criterial Features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.

Jarvis, S., Castaneda-Jiménez, G., and Nielsen, R. (2012). Detecting l2 writers' l1s on the basis of their lexical styles. In Jarvis, S. and Crossley, S. A., editors, *Approaching language transfer through text classification: Explorations in the detection-based approach*, pages 34–70.

Klieme, E. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie*. Verlagsgruppe Beltz.

Lüdeling, A., Walter, M., Kroymann, E., and Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*.

Meurers, D. (2013). Natural language processing and language learning. In Chapelle, C. A., editor, *Encyclopedia of Applied Linguistics*, pages 1–13. Blackwell.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Nesselhauf, N. (2005). *Collocations in a learner corpus*, volume 14. John Benjamins.

Nussbaumer, M. and Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In Sieber, P., editor, *Sprachfähigkeiten–besser als ihr Ruf und nötiger denn je*, pages 141–186. Sauerländer.

O'Donnell, M. B. and Römer, U. (2012). From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 7(1).

Reznicek, M., Ludeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., and Andreas, T. (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Technical report, Department of German Studies and Linguistics, Humboldt University, Berlin, Germany.

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013).

Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In Díaz-Negrillo, A., Ballier, N., and Thompson, P., editors, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123. John Benjamins.

Römer, U. and O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

Spinelli, B. and Parizzi, F., editors. (2010). *Profilo della lingua italiana*. La Nuova Italia.

Thelen, T. (2010). *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Ph.D. thesis, Dissertation.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 163–173, Montreal, Canada. Association for Computational Linguistics.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of corpus linguistics*, volume 9.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). IDS: Grammatik der Deutschen Sprache.

Zinsmeister, H. and Breckle, M. (2012). The ALeSKo learner corpus. *Multilingual Corpora and Multilingual Corpus Analysis*, 14:71–96.

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., and Skiba, D. (2012). Das Wissenschaftliche Netzwerk "Kobalt-DaF". *Zeitschrift für Germanistische Linguistik*, 40(3):457–458.

Zipser, F. and Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*.