Anna Corazza, Simonetta Montemagni and Giovanni Semeraro (dir.)

Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016
5-6 December 2016, Napoli

Accademia University Press

An extended version of the KoKo German L1 Learner corpus

Andrea Abel, Aivars Glaznieks, Lionel Nicolas and Egon Stemle

DOI: 10.4000/books.aaccademia.1743
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2016
Published on OpenEdition Books: 26 July 2017
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899982546

An extended version of the KoKo German L1 Learner corpus

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

Abstract

English. This paper describes an extended version of the KoKo corpus (version KoKo4, Dec 2015), a corpus of written German L1 learner texts from three different German-speaking regions in three different countries. The KoKo corpus is richly annotated with learner language features on different linguistic levels such as errors or other linguistic characteristics that are not deficit-oriented, and is enriched with a wide range of metadata. This paper complements a previous publication (Abel et al., 2014a) and reports on new textual metadata and lexical annotations and on the methods adopted for their manual annotation and linguistic analyses. It also briefly introduces some linguistic findings that have been derived from the corpus.

Italiano. *Il contributo descrive una versione estesa del corpus KoKo (versione KoKo4, Dic 2015), corpus che raccoglie produzioni scritte di apprendenti di tedesco L1, provenienti da tre distinte regioni germanofone, a loro volta situate in tre diversi paesi. Il corpus KoKo è annotato dettagliatamente su differenti livelli linguistici rilevanti, quali gli errori o altre caratteristiche linguistiche non direttamente ricollegabili a deficit individuali, ed arricchito da un'ampia gamma di metadati. Questo contributo integra una precedente pubblicazione (Abel et al., 2014a) è informa sui nuovi metadati testuali e sulle nuove annotazioni lessicali cosi come sui metodi adottati per la loro annotazione manuale e per le loro analisi linguistiche. Inoltre presenta brevemente alcuni risultati ricavati dal corpus.*

1 Introduction

The study of linguistically annotated learner corpora has received a growing interest over the past 20 years (Granger et al., 2013). In learner corpus linguistics, such corpora are usually defined as "systematic computerized collections of texts produced by language learners" (Nesselhauf, 2005). Unlike most learner corpora focusing on L2/FL learners (i.e. learners learning a foreign language), the KoKo corpus focuses on advanced L1 speakers that are still learning their mother tongue, which typically happens in educational contexts.

This paper describes an extended version of the KoKo corpus (Abel et al., 2014a), a corpus created for the purposes of the KoKo project which aims at investigating the writing skills of German-speaking secondary school pupils. The creation of the corpus was guided by two goals: on the one hand to describe writing skills at the end of secondary school, on the other hand to consider external socio-linguistic factors (e.g. gender, socio-economic background etc.).

The previous description focused on the data collection, the data processing, the annotation of orthographic and grammatical features as well as on aspects regarding annotation quality (Abel et al., 2014a). This paper, however, introduces the new textual metadata and lexical annotations.

The paper is structured as follows. In section 2, key facts are briefly reported, including references to related work. The new textual metadata and lexical annotations are then described in section 3, alongside with the methods adopted for their manual annotation and linguistic analyses and some examples of linguistic findings. In section 4, future works are discussed right before concluding in section 5.

2 Key Information about the Corpus

The KoKo corpus is a collection of 1,503 authentic argumentative essays, and the corresponding survey information about their authors, produced in classrooms under standardized conditions by learners of 85 classes of 66 schools from three different German-speaking areas: South Tyrol in Italy, North Tyrol in Austria and Thuringia in Germany.[1] Such areas are particularly suitable for comparative studies because of differences regarding the German standard varieties, the use of dialectal vs. standard varieties and the monolingual vs. plurilingual environments (Abel et al., 2014a).

The corpus is roughly equally distributed over the three regions and amounts to 824,757 tokens (punctuation excluded). All writers were attending secondary schools one year before their school-leaving examinations. 83% of the pupils were native speakers of German. The corresponding L1 part of the corpus amounts to 726,247 tokens. Metadata annotations amount to 52,605 annotations whereas manual annotations amount to 117,422 annotations. Furthermore, 366 features to measure linguistic complexity[2](Hancke et al., 2012; Hancke and Meurers, 2013) were automatically calculated per text (550,098 in total) and added as metadata.

Previous evaluation showed high accuracy of manual transcriptions ($> 99\%$), and automatic tokenization ($> 99\%$), sentence splitting ($> 96\%$) and POS-tagging ($> 96\%$) (Glaznieks et al., 2014).

As it is among the first accessible richly linguistically annotated German L1 learner corpora, the KoKo corpus is particularly relevant to L1 learner language researchers, and for the field of didactics of German as L1. Other comparable language resources are either not accessible (Berg et al., 2010; DESI-Konsortium, 2006; Nussbaumer and Sieber, 1994), or although accessible, have not been enriched with linguistic information (Augst et al., 2007; Fix and Melenk, 2002) or are only partly annotated (Thelen, 2010). Some other corpora include L1 data, but as reference for L2/FL learner corpus research (Reznicek et al., 2010; Zinsmeister and Breckle, 2012).

[1] We followed the privacy policy for such surveys and requested a signed consent from all adult participants and parents of minors. In addition, all students participated anonymously, no names of the students were collected, names of schools were codified and made anonymous.

[2] e.g. syntactic features such as the average length of NPs, VPs and PPs as well as their number per sentence, morphological features such as the number of modal verbs per total number of verbs or the average compound depth of nouns, and lexical features such as lexical diversity described by means of different measures

3 New Metadata and Annotations

This section describes the main features of the latest corpus version KoKo4 (Dec. 2015) that have been added to the version KoKo3 (Dec. 2014). It thus focuses on a new set of textual metadata and a new layer of lexical annotations which is, due to the selected features and the degree of granularity, a novelty in (corpus-based) modeling of L1-writing competences for German .

3.1 Textual Metadata

In the KoKo corpus, two kinds of **Metadata information** are available: (1) non-linguistic, i.e. person-related information provided by each participant via a questionnaire survey in class that is available for the whole sample and (2) linguistic, i.e. text-related information provided for a subsample of the corpus (569 texts, equally distributed over the three regions involved) through an online evaluation form by three different specially trained raters originating from the different participating regions.

While type (1) metadata allow for sociolinguistic analyses in order to detect relations between linguistic features (e.g. text length, sentence length, orthographic errors, grammatical errors, etc.) and non-linguistic person-related information, type (2) metadata constitute a further expansion of our analysis by including textual features as well. Text analysis was done holistically using an evaluation form and detailed guidelines that were elaborated on the basis of recent findings in writing research and text analyses (Brinker, 2010; Feilke, 2010; Augst et al., 2007; Böttcher and Becker-Mrotzek, 2006; Jechle, 1992; Augst and Faigel, 1986) and the curricula in the participating regions. The text evaluation form distinguishes four categories : **(A) formal completeness, (B) content, (C) formal and linguistic means of text arrangement and (D) overall impression**.

For **category A**, 10 questions of the online evaluation form focused on the presence of obligatory text parts (introduction, main part, closing part) and explicitly requested constituents of argumentative essays (opinion of the author, conclusion). The 25 questions of **category B** belong to

two subcategories: (B1) the topics of the essay (9 questions), (B2) patterns of topic development (16 questions). B1 comprises evaluations on e.g. the topics of each text part, gaps, and the overall coherence of the text. B2 refers to the main pattern of topic development (argumentative, etc.), the argumentation strategies (point of view, concessive or not), and the motivation of arguments (objective vs. subjective stance, quality of arguments). Formal and linguistic means of text arrangement (**category C**, 7 questions) focus on the use of paragraphs, the explicit announcement of and commitment to the function of the essay, and the use of linguistic means to structure the text with regards to content. Finally, **category D** (20 questions) aims for an overall impression and therefore focuses on the completion of the task (successful or not), the overall quality of the text and the overall consistency of both the quality and coherence. Of all 62 questions of the entire online evaluation form, we used 57 for each document of the subcorpus (alltogether 33,972 annotations).

The analyses revealed, among other things, that the text quality is classified as quite satisfactory on a 5 point Likert-scale[3]. More specifically, there are significant correlations between text quality assessment and other linguistic variables: thus, a lower number of e.g. lexical errors is connected to a higher text quality score[4], and, finally, a variety of group differences could be detected (e.g. concerning school type: lower text quality scores within vocational schools compared to general high schools[5]).

3.2 Lexical Annotations

As for the **manual annotations** of orthographic and grammatical features added to previous corpus versions (Abel et al., 2014a), a specifically crafted tag set and annotation manual were used for the annotation of lexical features. 61,728 lexical annotations were manually performed by trained annotators on a subcorpus of 980 texts, almost equally distributed over the three regions.

The analyses of lexical features focuses on lexical knowledge as a central part of lexical competence which includes the dimensions of lexical breadth (quantitative aspect) and lexical depth (qualitative aspect) (Steinhoff, 2009; Böttcher and Becker-Mrotzek, 2006; Mukherjee, 2005; Read and Nation, 2004; Read, 2000; Nation, 2001). Whereas the analyses of quantitative aspects of lexical knowledge were performed automatically by using different measures (e.g. lexical diversity measures such as MTDL and Yule's K, or lexical frequency scores based on dlexDB (Hancke and Meurers, 2013)), the analyses of qualitative aspects were done by means of manual annotations. We focus hereafter exclusively on the manual annotations allowing us to model qualitative aspects of lexical knowledge.

Category	Sub-category	Total
Single words	Neol. & occas.	4,670
	Arg. adv. & conj.	14,345
Phrasemes	Referential	18,708
	Communicative	4,824
	Structural	2,704
Particularities	Semantic	8,397
	Stylistic	236
	Form	1,923
	Metalinguistic	1,412
Target hyp.		4,509

Table 1: Quantitative figures for the 980 documents annotated with the new lexical annotations.

For annotating lexical features, we developed a new hierarchically-structured linguistic classification scheme inspired by previous work that focused on L2 learner languages (Abel et al., 2014b; Konecny et al., 2016). The classification scheme takes both into account occurrences of selected lexical phenomena and defective as well as non-defective particularities of learner languages considering two dimensions: (1) the linguistic subcategory, e.g. collocations and idioms, and (2) a target modification classification, e.g. omission, addition (Díaz-Negrillo and Domínguez, 2006; Abel et al., 2014b). Furthermore, we formulated target hypotheses for those categories that we annotated as defective in order to make the error interpretation transparent (Lüdeling et al., 2005). The corresponding annotation scheme contains 77 different tags including a set of further attributes.

In a multi-stage annotation procedure, all occurrences of phenomena on both single words and formulaic sequences (FS) were annotated (Wray, 2005). Annotations for particularities were subsequently added in order to distinguish between er-

[3] percentages: 1 (scarse): 6.2 - 2: 22.9 - 3: 39.0 - 4: 26.3 - 5 (excellent): 5.7)

[4] Kruskal Wallis H Test: FS errors $X^2(1) = 10.417$, $p = .036$, single word errors: ANOVA $F(4, 338) = 2.805$, $p = .026$

[5] Kruskal Wallis H Test: $X^2(1) = 49.147$, $p = .000$

rors concerning correctness, errors concerning appropriateness of usage (Eisenberg, 2007; Schneider, 2013), non-defective modifications (to capture, for example, creative use of language), and diasystematic markedness. At the single word level, we considered all out-of-vocabulary tokens of the part-of-speech tagger (Schmid, 1994) as candidates of neologisms or occasionalisms. In addition, we captured a variety of tokens relevant for the text genre of an argumentative essay (i.e. argumentative adverbs and conjunctions). At the level of FS, we applied a function-based approach distinguishing between three main categories of phrasemes (Burger, 2007), each of them with further subcategories (Abel et al., 2014b; Konecny et al., 2016; Granger and Paquot, 2008; Burger, 2007; Stein, 2007; Steinhoff, 2007), as well as a "mixed classification" (Burger, 2007):

Referential phrasemes include collocations[6] and idioms[7], distinguished among other things with respect to their degree of idiomaticity. **Communicative phrasemes** are subdivided into those bound to specific situations[8], and those not bound to specific situations[9]. Finally, **structural phrasemes** comprise complex conjunctions and prepositions[10] and concessive constructions[11].

For particularities, we considered four main categories, each with further subcategories:

On a **semantic dimension** a distinction is made between denotative errors concerning correctness or appropriatness of use[12], and connotative markedness or appropriatness of use[13]. The **stytlistic dimension** considers repetition, and redundancy. The **form dimension** focusses on word formation errors (concerning single word units only[14]), and on omission, choice, position and addition errors as well as creative modifications (concerning FS). Concerning **metalinguistic markers** the appropriateness of the use of quotation marks for highlighting units is considered.

An overview of the number of annotations is provided in Table 1.

Results of the analyses showed, among others, that pupils use different types of FS quite frequently, on average 5.12 constructions per 100 words: with 62%, non idiomatic referential phrasemes constitute the major part, followed by idiomatic referential phrasemes (19%), and, finally, structural (10%) and communicative phrasemes (9%). However, lexical errors in general affect more often FS than single word units (10% of the FS vs. 1.04% of the single words). The latter are most frequently form errors (5.50% of FS affected, especially choice errors: 4.17%).

4 Future Work

The KoKo project was completed and presented to the public in December 2015. We will start releasing the data via the corpus exploration interface ANNIS3 (Krause and Zeldes, 2016) and for download on request, after signing a license agreement.[15] Aside from the aforementioned data, future versions will also include additional metadata information about the authors integrated for the purposes of future socio-linguistic analyses.

Consensus in the annotations among annotators, and as such an indication of its reliability, will be evaluated on sub-sets of texts that were annotated for this purpose by more than one annotator. Three annotators independently annotated the text level metadata annotations on 27 texts, and six annotators independently annotated the lexical level annotations on the same 27 texts. Inter-annotator agreement will be calculated for annotations and segmentation, i.e. the agreement on the decision which word sequence needs to be tagged vs. what annotation needs to be assigned to it, and will be evaluated and reported in the form of Fleiss Multi-k and boundary similarity (Artstein and Poesio, 2008; Fournier, 2013).

Finally, thanks to its relatively large size and its richly annotated nature, potential additional uses

[6] further divided into restricted and loose collocations, light verb constructions (called "Funktionsverbgefüge" in German) as well as special classes such as irreversible bi- and trinominals, similes etc.

[7] further divided into nominative idioms and fixed phrases, and special classes such as irreversible bi- and trinominals etc.

[8] further divided into general routine or speech act formulas, special classes such as commonplaces, slogans, proverbs etc., and empty formulas

[9] further divided into text organising formulas, and interaction organising formulas

[10] further divided into phraseological connectors and syntactically complex connectors, and secondary prepositions

[11] further divided into constructions with "although" and a correlate of the "but"-class, and constructions with a modal word and a correlate of the "but"-class

[12] further divided into reference/function, contextual fitness, semantic compatibility, and precision

[13] further divided into speaker's attitude, and diasystematic markedness concerning language usage, i.e. diaphasic markedness, diachronic markedness, diatopic markedness

[14] distinguishing between errors with respect to derivation and to composition

[15] We have been trying to make the data available for direct download – but have to take more legal hurdles.

of the KoKo corpus in Natural Language Processing and Corpus Linguistics are being considered. Regarding Natural Language Processing, the error annotations paired with target hypothesis annotations allow for creating an aligned corpus. Such corpora can be used to improve machine translation for automatically correcting learner texts (Ng et al., 2014). Regarding Corpus Linguistics, machine learning methods can be used (e.g. as being done in WebAnno (Yimam et al., 2014)) to drive linguistic intuitions when performing annotations or analyses. Because of the richness of its annotation schemes, the KoKo corpus constitutes a challenging but at the same time promising dataset to test if the developed methods are able to uncover relevant correlations that have already been investigated, or to uncover even new ones that are worth considering for future linguistic analyses.

5 Conclusion

This paper described the most recent version of the KoKo corpus, a collection of richly annotated German L1 learner texts, and focused on the new textual metadata and lexical annotations.

Because other comparable language resources are either not accessible, or have not been enriched with linguistic information or are only partly annotated, the corpus is a valuable resource 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.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014a. Koko: An L1 learner corpus for german. In *Proceedings of LREC 2014*, pages 2414–2421.

Andrea Abel, Katrin Wisniewski, Lionel Nicolas, and Detmar Meurers. 2014b. A trilingual learner corpus illustrating european reference levels. *RICOGNIZIONI— Rivista di Lingue, Letterature e Culture Moderne*, 1(2):111–126.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Gerhard Augst and Peter Faigel. 1986. *Von der Reihung zur Gestaltung: Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13-23 Jahren*, volume 5 of *Theorie und Vermittlung der Sprache*. Peter Lang, Frankfurt.

Gerhard Augst, Katrin Disselhoff, Alexandra Henrich, Thorsten Pohl, and Paul Völzing. 2007. *Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter*. Peter Lang, Frankfurt.

Margit Berg, Anne Berkemeier, Reinold Funke, Christian Glück, Christiane Hofbauer, and Jordana Schneider, editors. 2010. *Sprachliche Heterogenität in der Sprachheil- und der Regelschule. Abschlussbericht im Programm „Bildungsforschung" der Landesstiftung Baden-Wrttemberg, Germany*.

Ingrid Böttcher and Michael Becker-Mrotzek. 2006. *Schreibkompetenz entwickeln und beurteilen*. Cornelsen, Berlin.

Klaus Brinker. 2010. *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. Bearbeitet von Sandra Ausborn-Brinker, 7., durchgesehene Auflage*, volume 29 of *Grundlagen der Germanistik*. Erich Schmidt Verlag, Berlin.

Harald Burger. 2007. *Phraseologie: Eine Einführung am Beispiel des Deutschen*, volume 36 of *Grundlagen der Germanistik*. Erich Schmidt Verlag, Berlin.

DESI-Konsortium, editor. 2006. *Unterricht und Kompetenzerwerb in Deutsch und Englisch*. Beltz Verlag, Weinheim – Bern.

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

Peter Eisenberg. 2007. Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm. *Aptum. Zeitschrift für Sprachkritik und Sprachkultur*, 3(2007):209–228.

Helmuth Feilke. 2010. Schriftliches Argumentieren zwischen Nähe und Distanz am Beispiel wissenschaftlichen Schreibens. *Nähe und Distanz im Kontext variationslinguistischer Forschung*, pages 209–231.

Martin Fix and Hartmut Melenk. 2002. *Schreiben zu Texten-Schreiben zu Bildimpulsen: das Ludwigsburger Aufsatzkorpus; mit 2300 Schülertexten, Befragungsdaten und Bewertungen auf CD-ROM*. Schneider-Verlag, Hohengehren.

Chris Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the ACL*, pages 1702–1712. ACL.

Aivars Glaznieks, Lionel Nicolas, Egon Stemle, Andrea Abel, and Verena Lyding. 2014. Establishing a standardised procedure for building learner corpora. *Apples - Journal of Applied Language Studies*, 8(3):5–20.

Sylviane Granger and Magali Paquot. 2008. Disentangling the phraseological web. *Phraseology. An interdisciplinary perspective*, pages 27–50.

Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. 2013. *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research Conference (LCR 2013)*, pages 54–56.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 1063–1080, Mumbai.

Thomas Jechle. 1992. *Kommunikatives Schreiben: Prozess und Entwicklung aus der Sicht kognitiver Schreibforschung*, volume 41 of *ScriptOralia*. Gunter Narr Verlag, Tübingen.

Christine Konecny, Andrea Abel, Erica Autelli, and Lorenzo Zanasi. 2016. Identification and Classification of Phrasemes in an L2 Learner Corpus of Italian. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology*, pages 533–542. Editions Tradulex, Geneva.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, pages 15–17.

Joybrato Mukherjee. 2005. The native speaker is alive and kicking: Linguistic and language-pedagogical perspectives. *Anglistik*, 16(2):7–23.

I.S.P. Nation. 2001. *Learning Vocabulary in Another Language*. Foreign Language Study. Cambridge University Press.

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus*, volume 14 of *Studies in Corpus Linguistics*. John Benjamins Publishing, Amsterdam.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. ACL.

Markus Nussbaumer and Peter Sieber. 1994. Texte analysieren mit dem Zürcher Textanalyseraster. In Peter Sieber, editor, *Sprachfähigkeiten–Besser als ihr Ruf und nötiger den je!*, pages 141–186. Verlag Sauerländer, Aarau.

John Read and Paul Nation. 2004. Measurement of formulaic sequences. In Norbert Schmitt, editor, *Formulaic sequences: Acquisition, processing and use*, Language Learning & Language Teaching, pages 23–35. John Benjamins Publishing, Amsterdam.

John Read. 2000. *Assessing vocabulary*. Cambridge University Press.

Marc Reznicek, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas. 2010. Das Falko-Handbuch. Korpusaufbau und Annotationen. Technical report, Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Jan Georg Schneider. 2013. Sprachliche ‚Fehler' aus sprachwissenschaftlicher Sicht. In *Sprachreport*, volume 1-2/2013, pages 30–37. Institut fr Deutsche Sprache, Mannheim.

Stephan Stein. 2007. Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In Harald Burger, Dmitrij Dobrovolskij, Peter Kühn, and Neal R. Norrick, editors, *Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung*, volume 1, pages 220–236. de Gruyter, Berlin – New York.

Torsten Steinhoff. 2007. *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*, volume 280 of *Reihe Germanistische Linguistik*. de Gruyter, Berlin – New York.

Torsten Steinhoff. 2009. *Wortschatz–eine Schaltstelle für den schulischen Spracherwerb?*, volume 17/2009 of *SPASS*. Universität Siegen, FB 3.

Tobias Thelen. 2010. *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Ph.D. thesis, University of Osnabrück.

Alison Wray. 2005. *Formulaic language and the lexicon*. Cambridge University Press.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In Kalina Bontcheva and Zhu Jingbo, editors, *Proceedings of the 52nd Annual Meeting of the ACL. System Demonstrations*, pages 91–96. ACL, jun.

Heike Zinsmeister and Margit Breckle. 2012. The alesko learner corpus: design–annotation–quantitative analyses. *Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: John Benjamins*, pages 71–96.