A data-driven platform for creating educational content in language learning*

Konstantin Schulz, Andrea Beyer, Malte Dreyer, and Stefan Kipf

Humboldt-Universität zu Berlin, Germany

Abstract. In times of increasingly personalized educational content, designing a data-driven platform which offers the opportunity to create content for different use cases is arguably the only solution to handle the massive amount of information. Therefore, we developed the software "Machina Callida" (MC) in our project CALLIDUS (Computer-Aided Language Learning: Vocabulary Acquisition in Latin using Corpus-based Methods).

The main focus of this research project is to optimize the vocabulary acquisition of Latin by using a data-driven language learning approach for creating exercises. To achieve that goal, we were facing problems concerning the quality of externally curated research data (e.g. annotated text corpora) while curating educational materials ourselves (e.g. predefined sequences of exercises). Besides, we needed to build a user-friendly interface for both teachers and students. While teachers would like to create an exercise or test and use them (even as printed out copies) in class, students would like to learn on the fly and right away.

As a result, we offer a repository, a file exporter for various formats and, above all, interactive exercises so that learners are actively engaged in the learning process. In this paper we show the workflow of our software and explain the architecture focusing on the integration of Artificial Intelligence (AI) and data curation. Ideally, we want to use AI technology to facilitate the process and increase the quality of content creation, dissemination and personalization for our end users.

Keywords: Educational content · Language learning · Data-driven · Exercise repository

1 Curating language exercises: the user's point of view

In German high schools, Latin is to this day the third most important foreign language, esp. in grades 7 to 10. For that reason, educational publishing companies are investing in teaching materials for Latin classes, but all these materials bear certain challenges for educational stakeholders: they are proprietary, hardly

[*] This project is funded by the German Research Foundation (project number 316618374) and lead by Malte Dreyer, Stefan Kipf and Anke Lüdeling. Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

adaptable (or not even digital) for teachers and split into a vast amount of different items like textbook, exercise book, vocabulary book etc. that all learners have to buy separately, if needed [7, p. 194f.]. On top of that, most of the teaching materials only refer to the initial stage of language acquisition, in which Latin original texts do not yet matter [21, p. 133]. Although the companies are also providing teachers with reading books for intermediate learners containing sections of selected Latin original texts, teachers are still in continuous need of adaptable texts and exercises for these advanced stages. In addition, although the curricula offer a standardized canon of Latin authors [20, p. 45], it still includes a wide range of different texts compared to the available time in Latin classes. What is more, teachers prefer to use texts whose vocabulary is covered as much as possible by the basic vocabulary already acquired by the students, since the comprehensibility of the text can be considerably limited if less than 95% of words are known [25, p. 352].

As a consequence, teachers may choose texts from a large pool of Latin authors, but without supporting material they rarely do, because they lack the time to prepare texts and exercises independently. Instead, they often fall back on ready-made materials that are quality-tested but rarely fit the needs of the learning group. This situation results in a kind of dilemma: Many teachers would like to enrich their lessons with further authors and support their students individually in their language acquisition with (personalized) exercises, but they do not feel up to the challenge of selecting and adapting materials to their students' needs [24, p. 115/117].

This brief outline of the problem shows the need to develop a platform that allows teachers (and students) to create needs-based exercises for authentic Latin texts. Furthermore, for a good user experience it is necessary that the process of generation is fast and easy to handle, that the generated exercises are ready to use (analogically and digitally) or share, and that they are well curated for later reuse. These requirements are illustrated in three exemplary use cases which have been modeled loosely following the guidelines of Cockburn [9].

Use Case	1: The teacher needs exercises based on authentic Latin texts	2: The teacher does not have enough time to prepare an exercise manually	3: The teacher wants to support his/her students in a personalized way to enable individual learning
Primary actor	Teacher		
Stakeholders	Teacher, students		
Scope	An easy to handle exercise generator	A database with well-curated different types of exercises	Learning Analytics and recommendations for future exercises

User story	As a teacher, I want to select a section of the work to be read. I want to compare this section to the used core vocabulary for getting an overview of the amount of unknown words. Then, I want to set the parameters of the intended exercise: type of exercise and linguistic focus (specific lemmata, syntactic structures, morphology, context-based meaning, word equivalents). After getting a preview, all selections can be easily changed, if I think that, e.g.,the exercise is too difficult.	As a teacher, I search the repository for at least one matching exercise. I want to combine different search terms in an extended search, e.g. Latin text passage, exercise type, linguistic focus, popular exercises, vocabulary. Then, I want to use the exercise in class (with smartphones, tablets or interactive whiteboard), to embed it in a learning platform for later use or to send it to the students for their homework.	As a teacher, I want an overview of how my students perform in an exercise. I want to be able to see at a glance what mistakes are made most often so that I know what to focus on when creating the next exercise. I would also like a recommendation as to which exercise to select next if there already is a suitable exercise in the database.
Level	Repetition and deepening of vocabulary knowledge in context	Repetition and deepening of vocabulary knowledge (individually)	Zone of proximal (linguistic) development of each student
Precondition	Teachers are presented with an option to generate new exercises.	Teachers are presented with an option to browse exercises from an existing database.	Students generate data about their individual progress. The data can be tracked and analyzed automatically.
Minimal Guarantees	The generated exercise can be exported.	The database contains exercises and can be searched.	Many students have completed the same (or similar) exercises.
Success Guarantees	The generated exercise can be shared and is stored in a database that is easily accessible to end users.	The search for a matching exercise is supported by advanced filtering. Popular and well-curated exercises are marked.	Teachers receive helpful suggestions for choosing the next exercise.

Trigger	The teacher invokes the exercise generation setup.	The teacher decides to use a ready-made exercise.	Students have just completed an exercise and now should attempt another one.
Basic flow: Step 1	The teacher picks the option of generating a new exercise.	The teacher picks the option of searching the database.	The students register with the software and go through the given exercise.
Step 2	The teacher chooses a text passage from a wide range of Latin authors.	The teacher selects a single or multiple filters or uses the extended search option.	The teacher receives an evaluation about the performance (percentage, error types) of each student.
Step 3	The teacher compares the words of the text with the used core vocabulary and changes the section accordingly (go to step 2) or proceeds to set the parameters of the exercise.	The teacher evaluates the results. Depending on the results, the teacher changes the search terms / filters (go to step 2) or decides to use one of the given exercises.	The teacher also gets a recommendation which parameters to set for the next exercise or which exercise to select from the database.
Step 4	The teacher decides on the exercise format, the linguistic focus and the instruction statement.	The teacher uses the exercise in class or disseminates it using a link, so that students may use their own mobile devices.	The students get their new exercise and work on it (go to step 1).
Step 5	The system presents a preview. The teacher either exports the exercise to a printable format or shares it digitally or tries other parameters (go to step 4) or even changes the section (go to step 2).		

Table 1: Use Cases

2 Automatic parsing and evaluation: the developer's point of view

In order to help teachers create high-quality educational content, we provide support for each of the necessary steps in our software at `https://korpling.org/mc`.

2.1 Selection of text (Use Case 1)

Many Latin text editions are proprietary and thus do not comply with the FAIR data principles [32]. Additionally, such resources are not compatible with the requirements for projects funded by the German Research Foundation, which need to prefer open licenses to closed ones [16]. To solve this problem, we decided to rely solely on text editions from the public domain. This choice also narrowed down the range of suitable text repositories a lot. In the end, we settled for the Perseus Library [3] because it has a well-defined API (Canonical Text Services [30]) and a standardized citation model (URN [8]) for ancient text passages, works and authors. This repository, however, offers a vast amount of texts: several hundreds of works from dozens of authors can be explored, so our users need a way to prioritize them according to their specific needs. Currently, we support this by offering a vocabulary filter and measures for text complexity.

The vocabulary filter has to be targeted at one of several reference vocabularies. These are essentially lemmatized word frequency lists derived from textbooks [5], treebanks [4] or materials created by publishing houses [31]. The reference vocabularies can be used to estimate the students' previous knowledge by specifying that, e.g., they should know the 500 most frequent words from that list. This subset of words is then compared to the lemmata occurring in a given corpus. Thus, if teachers specify a large corpus and the desired size of the final text passage, the software will rank all possible subsets of the corpus according to their congruence with the reference vocabulary. The boundaries for each subset are chosen intelligently in order to maximize the number of known words. This enables teachers to always choose a text that supports their students' zone of proximal development [27, p. 238].

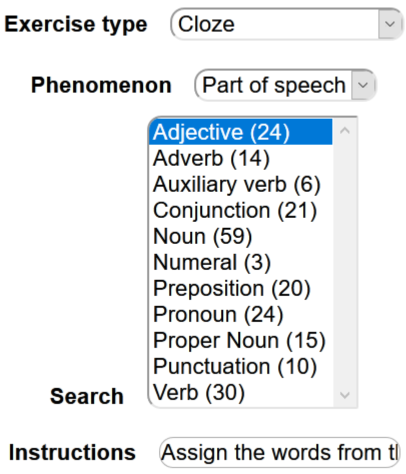
Text complexity, on the other hand, does not directly relate to a student's previous knowledge, but to an intrinsic comparison between multiple Latin texts. In our case, it is a combination of well-known operationalizations of the presumed degree of difficulty that readers may face when approaching a text, e.g. lexical density [19, p. 61]. This helps teachers to determine the suitability of a given text passage (or corpus) with regard to their students' linguistic competence. The major strength of such measures does not reside in their inherently flawed approximation of actual complexity, but in enabling a formalized linguistic comparison that goes beyond mere counting of words and integrates syntax, morphology and semantics [11, p. 607]. By combining information about vocabulary and text complexity, teachers can significantly accelerate and improve their choice of texts, thus curating better educational content for their students.

2.2 Focus on specific linguistic phenomena (Use Case 1)

Once teachers have committed themselves to a suitable text passage, they may still not know the exact target of a potential exercise. Therefore, we offer a keyword in context (KWIC) view to explore collocations and the specific usage of a particular word [18, p. 97]. The superficial token-based display is enriched by morpho-syntactic information, e.g. part of speech and dependency links. Therefore, teachers can qualitatively inspect usage patterns on multiple linguistic levels as needed.

A major problem in this approach is that most Latin texts are not curated as treebanks with scientific annotations, but rather just as plain text. In other words, we lack the key prerequisite to provide a rich KWIC view. To compensate for this shortcoming, we use an AI-driven dependency parser [29] to process plain Latin text in a fully automatic manner. It was trained as a multi-task classifier using representation learning on existing curated treebanks [28, p. 4291]. This is very reliable for basic tasks like tokenization, segmentation, lemmatization and part-of-speech tagging (>95% accuracy), but is rather error-prone (∼80% accuracy) for dependency links. Thus, the syntactic visualization in the KWIC view may not always be entirely correct, but the basic concordance function and the information about parts of speech are highly accurate, thereby enabling teachers to create educational content in a much more well-informed manner. Besides, the lack of performance on the syntactic level may be alleviated by accessing and linking further resources to the existing parser output [22, p. 75].

2.3 Design of interaction / learning setting (Use Case 1)



Exercise type: Cloze

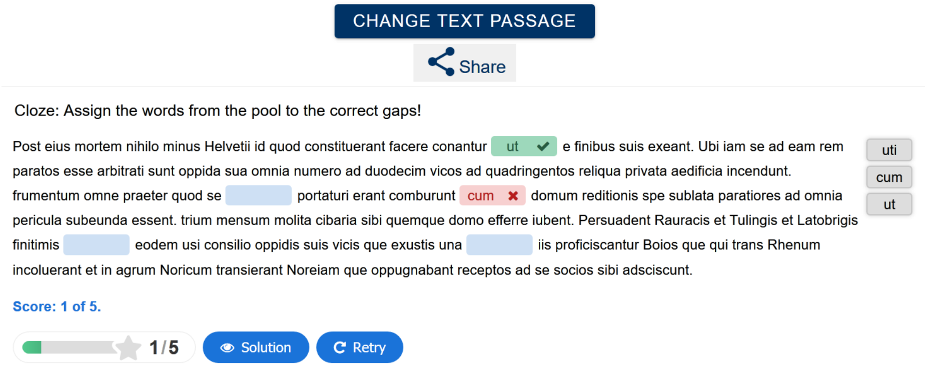
Phenomenon: Part of speech

Instructions: Assign the words from tl

Fig. 1. Setting parameters for a new exercise

Now that the basic content (i.e. texts and phenomena) of a new exercise has been established, it is time to look at the layout. Depending on the chosen phenomenon, but also on a student's personal preferences, certain types of interaction may be more appropriate than others in order to reach a specific educational goal (see Fig. 1). In general, a systematic variation of interaction types can support more learning styles [26, p. 169], make the learning process more multifaceted [17, p. 1] and lead to a higher degree of motivation [17, p. 5] and engagement [26, p. 165]. On the other hand, the exclusive usage of ready-made exercises in various formats can also cause mental overload for students [26, p. 161].

Therefore, we offer teachers the possibility to choose from a range of existing exercises with the same type of interaction, so it is easier for them to maintain a certain level of consistency, even in longer learning sequences. Furthermore, some of the exercise formats may be considered part of the same line of progression, e.g. clozes can be solved with a visible pool of boxes using Drag and Drop (easy, see Fig. 2) or by typing characters into blank text fields (more difficult). Besides, the same basic technology and layout can be used to produce different exercises, e.g. Drag and Drop works for both the cloze and matching format. In this regard, the usage of a large common framework (H5P [2]) allows for a diverse, but consistent learning experience. As an inspiration for longer sequences of exercises, we offer the so-called Vocabulary Unit which roughly corresponds to the length of an average lesson in school (about 45 minutes).



Cloze: Assign the words from the pool to the correct gaps!

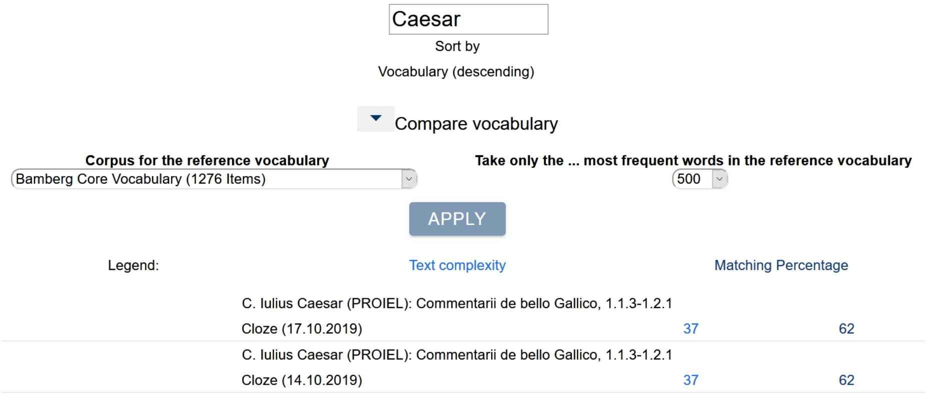
Post eius mortem nihilo minus Helvetii id quod constituerant facere conantur ut ✓ e finibus suis exeant. Ubi iam se ad eam rem paratos esse arbitrati sunt oppida sua omnia numero ad duodecim vicos ad quadringentos reliqua privata aedificia incendunt. frumentum omne praeter quod se portaturi erant comburunt cum ✗ domum reditionis spe sublata paratiores ad omnia pericula subeunda essent. trium mensum molita cibaria sibi quemque domo efferre iubent. Persuadent Rauracis et Tulingis et Latobrigis finitimis eodem usi consilio oppidis suis vicis que exustis una iis proficiscantur Boios que qui trans Rhenum incoluerant et in agrum Noricum transierant Noreiam que oppugnabant receptos ad se socios sibi adsciscunt.

Score: 1 of 5.

Fig. 2. Drag-and-Drop-based cloze exercise with visible pool and binary feedback

2.4 Dissemination (Use Case 2)

When teachers are satisfied with their created content, they typically want to distribute it to their students to employ it in a didactic context. To that end, every exercise is labeled with a unique identifier, so it can be saved in a database and shared via deep links to the software server (e.g. `https://korpling.org/mc/exercise?eid={EXERCISE_ID}`). When creating an exercise as well as at any later point in time, users may also export a given exercise to specific file formats: PDF and DOCX for printing, XML for integration into a learning management system. That way, teachers and students are able to build their own collections of useful exercises over time and, in the case of XML, derive additional benefit from the features offered by Learning Management Systems like Moodle [1]: structured online courses, user management, learning analytics and so on. If, on the other hand, teachers do not have the time to curate their own content, we provide access to public exercises that can be filtered and searched for using an extensive metadata schema, including the author, work, text passage, interaction type, popularity, vocabulary and text complexity (see Fig. 3).



Caesar

Sort by

Vocabulary (descending)

Compare vocabulary

Corpus for the reference vocabulary: Bamberg Core Vocabulary (1276 Items)

Take only the ... most frequent words in the reference vocabulary: 500

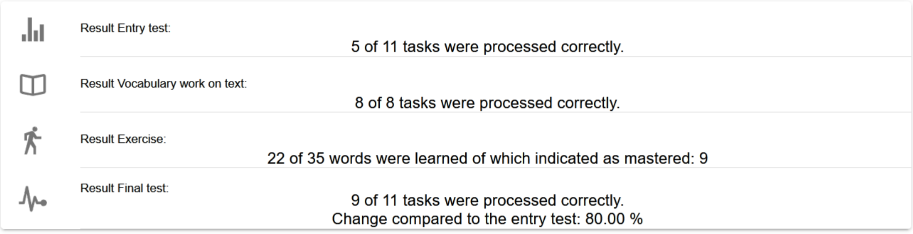
APPLY

Legend:	Text complexity	Matching Percentage
C. Iulius Caesar (PROIEL): Commentarii de bello Gallico, 1.1.3-1.2.1 Cloze (17.10.2019)	37	62
C. Iulius Caesar (PROIEL): Commentarii de bello Gallico, 1.1.3-1.2.1 Cloze (14.10.2019)	37	62

Fig. 3. Exercise Repository with keyword search and options for sorting/filtering

2.5 Evaluation (Use Case 3)

Moodle already offers summative evaluation for created exercises, but teachers usually refrain from using it because they have not been trained [6, p. 160] to deal with the technological complexity during setup, maintenance and everyday usage [10, p. 342]. This also applies to digital media in general [14, p. 18]. Therefore, in the long run, we need to provide such evaluation ourselves. A basic prototype that goes beyond the single-exercise binary feedback (correct/incorrect) has been implemented in our Vocabulary Unit. It shows the overall performance for the given exercises, the student's development from beginning to end and how many words from the target vocabulary are already known (see Fig. 4). In the future, we would like to add further analyses pertaining to the preferred type of interaction, problematic performance on certain linguistic phenomena and the speed of problem solving. These goals are in line with the recent trend of focusing on the learner's perspective in computer-assisted evaluation [15, p. 313]: Where are my strengths and weaknesses? How did I develop during the last weeks? What can I do to improve specific skills?



Result Entry test:
5 of 11 tasks were processed correctly.

Result Vocabulary work on text:
8 of 8 tasks were processed correctly.

Result Exercise:
22 of 35 words were learned of which indicated as mastered: 9

Result Final test:
9 of 11 tasks were processed correctly.
Change compared to the entry test: 80.00 %

Fig. 4. Summative evaluation of a student's performance in the Vocabulary Unit

However, user-specific quantitative evaluation is not enough. In order to increase students' learning success, they also need adaptive qualitative feedback. A prerequisite for that is the detection and classification of errors: the integrated binary evaluation of H5P can be used as a basis to categorize various error types, e.g.: Did the student fail to give any answer at all? Did the student actually provide the correct answer, but with minor typing mistakes? Did the student make obvious grammatical mistakes? If so, are they related to morphology, vocabulary or syntax? Depending on the specific type of error, suitable feedback needs to be generated. Our main objective here is to provide deeper support for teachers and students in order to optimize the learning progress towards a specific goal, e.g. being able to read texts from a specific corpus. A good approach in that case may be to create exercises for this corpus and use the students' performance as an objective for reinforcement learning [13, p. 2094]. The AI model should then learn to utilize suitable pedagogical actions (e.g. distributing exercises for learning) to maximize a student's performance on the test exercise dataset for a corpus.

3 Next steps: Learning Analytics and semantic analysis

For the future integration of Learning Analytics in our software, we have already built a prototype that evaluates a learner group's performance across multiple dimensions, e.g. working speed, interaction type, accuracy and performance gain over time. A large part of this analysis is most suitable for groups, which is why it is probably useful for teachers. Individuals, on the other side, would need a stronger emphasis on their development over time, which is harder to track because it would require them to use the software as their main source of language learning. Therefore, specific milestones are to be reached in the next months:

- summarize group performances as an indicator that helps teachers to readjust their general didactic strategy, e.g. by focusing more heavily on certain linguistic phenomena
- analyze results for individual students over time and suggest the most suitable exercises for them considering their personal characteristics, i.e. learning style, thematic priority and particular weaknesses

Apart from improving the quality of the existing workflow, we also consider increasing its quantity, e.g. by adding new linguistic phenomena: Semantics is currently underrepresented in our automatic analyses, which makes it hard for teachers to group their educational content around a certain topic. This could be alleviated by integrating representation learning as an independent feature: Unsupervised machine learning, in the form of Contextual Word Embeddings like those provided by BERT [12], may be used to distinguish different usages of the same word in different sentences, thereby highlighting fine-grained semantic differences between authors or even within the same work. While we already used Word2Vec [23] to perform simple vector-based analyses on existing Latin treebanks, it still remains a challenge to generalize the calculation, visualization and interpretation in this workflow while maintaining a sufficient level of quality. A well-founded evaluation of representation learning for the purposes of language acquisition is arguably the most important goal in this respect.

References

1. Moodle: A learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalised learning environments. Moodle Pty Ltd
2. H5P. Create, share and reuse interactive HTML5 content in your browser. Joubel AS (Jun 2018)
3. Almas, B., Babeu, A., Krohn, A.: Linked Data in the Perseus Digital Library. ISAW Papers **7**(3) (2014)
4. Bamman, D., Crane, G.: The Ancient Greek and Latin Dependency Treebanks [AGLDT]. In: Language Technology for Cultural Heritage, pp. 79–98. Springer (2011)
5. Bartoszek, V., Datené, V., Lösch, S., Mosebach-Kaufmann, I., Nagengast, G., Schöffel, C., Scholz, B., Schröttel, W.: VIVA 1 Lehrerband, vol. 1. Vandenhoeck & Ruprecht (2013)
6. Bäsler, S.A.: Lernen und Lehren mit Medien und über Medien. Ph.D. thesis, Technische Universität Berlin (2019). https://doi.org/http://dx.doi.org/10.14279/depositonce-7833
7. Beyer, A.: Das Lateinlehrbuch Aus Fachdidaktischer Perspektive: Theorie - Analyse - Konzeption. Universitätsverlag Winter GmbH, Heidelberg (2018)
8. Blackwell, C., Smith, N.: The Canonical Text Services URN Specification, Version 2.0.rc.1 [CITE / URN] (2015)
9. Cockburn, A.: Writing Effective Use Cases. The Agile Software Development Series, Addison-Wesley, Boston, 16. print edn. (2006)
10. Costa, C., Alvelos, H., Teixeira, L.: The use of Moodle e-learning platform: A study in a Portuguese University. Procedia Technology **5**, 334–343 (2012)
11. Dascalu, M.A., Gutu, G.S., Ruseti, S.S., Cristian Paraschiv, I.S., Dessus, P., Mcnamara, D.A., Crossley, S.A., Trausan-Matu, S.A.: ReaderBench: A Multi-lingual Framework for Analyzing Text Complexity. In: Lavoué, É., Drachsler, H., Verbert, K., Broisin, J., Pérez-Sanagustín, M. (eds.) Data Driven Approaches in Digital Education, Proc 12th European Conference on Technology Enhanced Learning, EC-TEL 2017. pp. 606–609. Data Driven Approaches in Digital Education 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings, Springer, Tallinn, Estonia (2017)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Dorça, F.A., Lima, L.V., Fernandes, M.A., Lopes, C.R.: Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis. Expert Systems with Applications **40**(6), 2092–2101 (May 2013). https://doi.org/10.1016/j.eswa.2012.10.014
14. Eickelmann, B., Bos, W., Labusch, A.: Die Studie ICILS 2018 im Überblick – Zentrale Ergebnisse und mögliche Entwicklungsperspektiven. In: Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., Vahrenhold, J., Eickelmann, B., Bos, W. (eds.) ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking, pp. 7–31. Waxmann (2019), oCLC: 1124310958

15. Ferguson, R.: Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning **4**(5/6), 304–317 (2012)
16. Forschungsgemeinschaft, D.: Appell zur Nutzung offener Lizenzen in der Wissenschaft. Tech. Rep. 68, Deutsche Forschungsgemeinschaft (Nov 2014)
17. Harecker, G., Lehner-Wieternik, A.: Computer-based Language Learning with Interactive Web Exercises. ICT for Language Learning pp. 1–5 (2011)
18. Helm, F.: Language and culture in an online context: What can learner diaries tell us about intercultural competence? Language and Intercultural Communication **9**(2), 91–104 (May 2009). https://doi.org/10.1080/14708470802140260
19. Johansson, V.: Lexical diversity and lexical density in speech and writing: A developmental perspective. Working Papers **53**, 61–79 (2008)
20. Kipf, S.: Geschichte des altsprachlichen Literaturunterrichts. In: Lütge, C. (ed.) Grundthemen Der Literaturwissenschaft, pp. 15–46. De Gruyter, Berlin and Boston (2019)
21. König, J.: Die Lektürephase. In: Janka, M. (ed.) Lateindidaktik, pp. 133–155. Cornelsen Scriptor, Berlin (2017)
22. Mambrini, F., Passarotti, M.: Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin. In: Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019). pp. 74–81. Paris, France (Aug 2019)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
24. Munser-Kiefer, M., Martschinke, S., Hartinger, A.: Subjektive Arbeitsbelastung von Lehrkräften in jahrgangsgemischten dritten und vierten Klassen. In: Miller, S., Holler-Nowitzki, B., Kottmann, B., Lesemann, S., Letmathe-Henkel, B., Meyer, N., Schroeder, R., Velten, K. (eds.) Profession und Disziplin : Grundschulpädagogik im Diskurs, pp. 114–120. Jahrbuch Grundschulforschung, Springer Fachmedien, Wiesbaden (2018)
25. Nation, I.S.: Learning Vocabulary in Another Language. Cambridge University Press, 2 edn. (2013)
26. Schmid, E.C.: Developing competencies for using the interactive whiteboard to implement communicative language teaching in the English as a Foreign Language classroom. Technology, Pedagogy and Education **19**(2), 159–172 (2010)
27. Shabani, K., Khatib, M., Ebadi, S.: Vygotsky's Zone of Proximal Development: Instructional Implications and Teachers' Professional Development. English language teaching **3**(4), 237–248 (2010)
28. Straka, M., Hajic, J., Straková, J.: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: LREC. pp. 4290–4297 (2016)
29. Straka, M., Straková, J.: UDPipe. A LINDAT/CLARIN project
30. Tiepmar, J., Teichmann, C., Heyer, G., Berti, M., Crane, G.: A new implementation for canonical text services [CTS]. In: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). pp. 1–8 (2014)
31. Utz, C.: Mutter Latein und unsere Schüler — Überlegungen zu Umfang und Aufbau des Wortschatzes [BWS]. Antike Literatur–Mensch, Sprache, Welt. Dialog Schule und Wissenschaft **34**, 146–172 (2000)
32. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S.,

Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018 (Mar 2016). https://doi.org/10.1038/sdata.2016.18