The Journal of Open Source Software

Hexatomic: An extensible, OS-independent platform for deep multi-layer linguistic annotation of corpora

Stephan Druskat [1,2*¶], **Thomas Krause** [3*], **Clara Lachenmaier** [2], and **Bastian Bunzeck**[2]

1 German Aerospace Center (DLR), Institute for Software Technology, Berlin, Germany **2** Friedrich Schiller University Jena, Department of English Studies, Jena, Germany **3** Humboldt-Universität zu Berlin, Department of German Studies and Linguistics, Berlin, Germany ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.04825

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @reckart

Submitted: 30 August 2022
Published: 01 June 2023

Summary

Linguistic research aims to understand how languages work. This includes finding regularities, understanding language use and/or production in a specific context, and other linguistic research topics. The data for this research are language artifacts, e.g., recordings or transcriptions of spoken language and records of written language, that are often collected in *corpora*. In order to analyze languages through these corpora, linguists apply different methods. Some of these methods are computational and include *automated annotation* of corpora through natural language processing, or gathering statistical insights into corpora through machine learning. Other methods are computer-aided, and include *manual annotation* of corpora across multiple annotation *layers*. Linguistic phenomena often occur on different aspects of language, e.g. morphology, part-of-speech, lemmatization, constituent and dependency syntax, entities, coreference, discourse and information structure. For a deep analysis and understanding of language data it is necessary to include annotation layers for these different aspects in a single multi-layer corpus. These annotations often have to be made manually, as linguists' insights into specific language features often surpass the generalized models used in machine learning. Additionally, for languages with fewer speakers such as those investigated in linguistic typology and language documentation projects, machine learning models may not exist or there may just not be enough language data for training model.

Hexatomic is an extensible platform for multi-layer linguistic annotation of corpora. It is available for Linux, MacOS and Windows systems. While it mainly targets manual annotation methods, it can be extended for automated annotation. *Hexatomic* merges previous architectural (Druskat et al., 2014) and functional (Gast et al., 2015) prototypes for such a platform. It uses a versatile graph-based model for linguistic data and includes converters between this model and different linguistic input and output formats (Zipser & Romary, 2010), to enable reuse and enrichment of existing corpora. The data model and the conversion framework, as well as the widely used corpus query and analysis platform ANNIS (Krause et al., 2022), are part of the corpus-tools.org family of linguistic software (Druskat et al., 2016), that *Hexatomic* completes.

Hexatomic enables users to build new corpora from scratch, or import existing corpora. Corpus projects allow flexible organization of corpora, sub-corpora and documents by storing them as nodes in a corpus graph. This way, many documents across many corpora can be maintained within a single project. These documents can then be annotated on arbitrary layers using different editor plugins. Hexatomic 1.0.1 includes a spreadsheet-like editor for token and span-based annotations, and a graph editor for arbitrary annotation layers. Additional annotation editors can be added to the platform as plugins, e.g., to perform bespoke, project-specific annotation tasks. Multiple editors can be used to simultaneously annotate the same data

(Figure 1), and any changes that are made update any editors currently showing the same elements, e.g., tokens. Annotation layers can also be filtered dynamically or manually for better usability.

Hexatomic comes with extensive user and developer/maintainer documentation, and can be automatically updated at runtime. The platform is built on the extensible Eclipse RCP platform for Java.

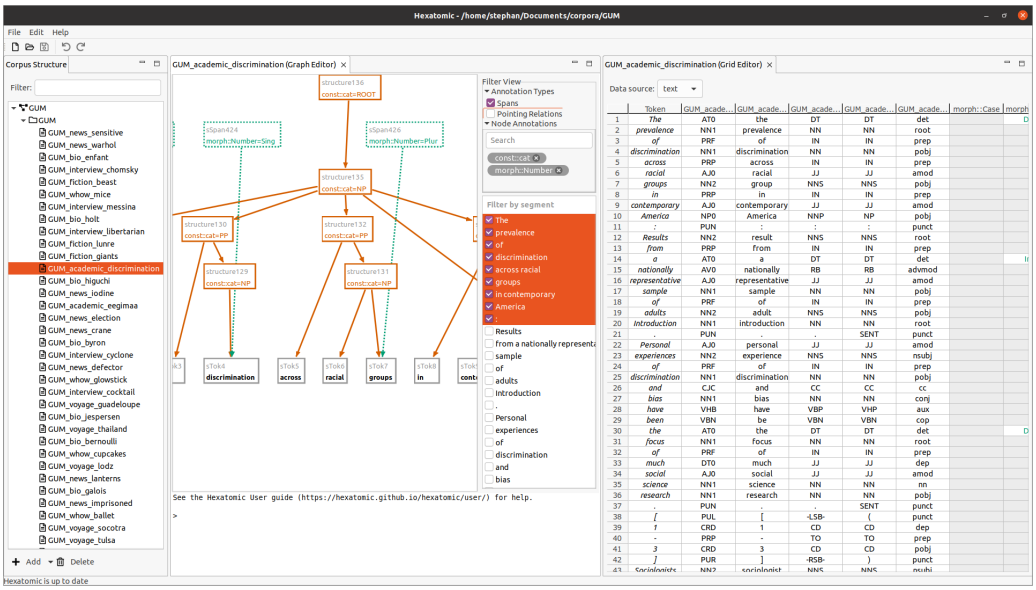

Figure 1: A screenshot of *Hexatomic* showing simultaneous annotation in graph and grid editors.

Statement of need

While there is a wide variety of corpus annotation tools available, most are "specialized single-user tools" (Biemann et al., 2017). *Hexatomic* meets a demand across different linguistic fields for an interdisciplinarily usable, highly compatible platform for multi-layer annotation of linguistic corpora. This is achieved through $n : m$ conversion capabilities, the generic nature of the data model, the wide applicability of the core editors, and the possibility to extend *Hexatomic* with new editors for specific annotation tasks. Corpora from different sources, e.g., from corpus linguistic and language documentation research, can be merged into a single project. Subsequently, they can be annotated in *Hexatomic* to help answer research questions from different linguistic disciplines, and finally exported to ANNIS for multi-layer search and analysis.

Web-based annotation tools such as WebAnno (Eckart de Castilho et al., 2016), INCEpTION (Klie et al., 2018) or GATE (Bontcheva et al., 2013) often run on centralized servers for easy access by users without the need for an installation procedure, even if some of these tools also support installation on end-user machines. Operating a server not only creates operating costs, but also complicates data transfer and constitutes a single point of failure, increasing the risk of data loss. Additionally, web-based annotation is impractical in regions without easily accessible internet connectivity, e.g., during linguistic fieldwork. *Hexatomic* is used on the researcher's local machine and offers full control over data. The local project folder holding the annotation data can be versioned with existing version control systems. By using Salt XML (Zipser & Romary, 2010) to store the project data and having separate files for each corpus document, conflicts in version control systems are minimized. Version controlled local corpus data can additionally be shared and collaborated on via existing collaboration platforms, e.g., those based on `git`.

Hexatomic has been designed and developed to gain a maximal potential for software sustainability. It is built on widely-used, mature technology with a strong community in research and industry, and explicitly designed for extensibility, adaption, and reuse. There is extensive documentation, including documentation for maintenance processes, changes in maintenance, and revival after periods without maintenance. *Hexatomic* will also be supported in the future through a long-term software maintenance and research software engineering position at the corpus linguistics working group at Humboldt-Universität zu Berlin.

Future work

Developed as a platform, *Hexatomic* will be used for annotating the RIDGES Herbology corpus (Odebrecht et al., 2017), a diachronic corpus of historic herbal texts. It is also planned to use it for the various corpora of the Deutsch Diachron Digital projects[1] that create a reference corpus of historical German texts for different time periods. To support these annotation projects along with other use cases, next development steps include:

- A specialized editor for different aligned tokenization layers of the same texts (Krause et al., 2012).
- Adding support for playing aligned audio and video files, which can already be linked and aligned in the data model. This will make Hexatomic applicable to multi-modal and multi-layer corpora such as RUEG (Wiese et al., 2020).

Through its current and new features, Hexatomic can greatly simplify complex linguistic annotation workflows, through avoiding the arduous and error-prone process of using different tools for different annotations and merging subsequently.

Acknowledgements

Hexatomic has been developed in the research project "A minimal infrastructure for the sustainable provision of extensible multi-layer annotation software for linguistic corpora". The project was funded under Deutsche Forschungsgemeinschaft's call "Research Software Sustainability", grant number 391160252, and ran from October 2018 until December 2021. Thomas Krause was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

References

Biemann, C., Bontcheva, K., Eckart de Castilho, R., Gurevych, I., & Yimam, S. M. (2017). Collaborative Web-Based Tools for Multi-layer Text Annotation. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 229–256). Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_8

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., & Gorrell, G. (2013). GATE Teamware: A web-based, collaborative text annotation framework. *Language Resources and Evaluation*, *47*(4), 1007–1029. https://doi.org/10.1007/s10579-013-9215-6

Druskat, S., Bierkandt, L., Gast, V., Rzymski, C., & Zipser, F. (2014). Atomic: An open-source software platform for multi-level corpus annotation. In J. Ruppenhofer & G. Faaß (Eds.), *Proceedings of the 12th edition of the konvens conference, hildesheim, germany, october 8-10, 2014* (pp. 228–234). Universitätsbibliothek Hildesheim.

Druskat, S., Gast, V., Krause, T., & Zipser, F. (2016). Corpus-tools.org: An interoperable generic software tool set for multi-layer linguistic corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J.

[1] http://www.deutschdiachrondigital.de/

Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation LREC 2016, portorož, slovenia, may 23-28, 2016*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/918.html

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. https://www.aclweb.org/anthology/W16-4011

Gast, V., Bierkandt, L., & Rzymski, C. (2015, April). Annotating modals with GraphAnno, a configurable lightweight tool for multi-level annotation. *Proceedings of the Workshop on Models for Modality Annotation*. https://aclanthology.org/W15-0303

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. https://www.aclweb.org/anthology/C18-2002

Krause, T., Benjamin, W., Rütte, T., Glushanok, I., Klotz, M., Zhang, S., Zeldes, A., Bartels, F., Druskat, S., Boyd, A., Stemle, E., Lampen, L., & Petran, F. (2022). *ANNIS* (Version 4.9.5). https://doi.org/10.5281/zenodo.1212548

Krause, T., Lüdeling, A., Odebrecht, C., & Zeldes, A. (2012). Multiple tokenization in a diachronic corpus. *Exploring Ancient Languages Through Corpora Conference (EALC)*. http://www.hf.uio.no/ifikk/english/research/projects/proiel/ealc/abstracts/Krause_et_al.pdf

Odebrecht, C., Belz, M., Zeldes, A., Lüdeling, A., & Krause, T. (2017). RIDGES herbology: Designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, *51*(3), 695–725. https://doi.org/10.1007/s10579-016-9374-3

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., Jahns, E., Klotz, M., Krause, T., Labrenz, A., Lüdeling, A., Martynova, M., Neuhaus, K., Pashkova, T., Rizou, V., Rosemarie, T., Schroeder, C., Szucsich, L., Tsehaye, W., … Zuban, Y. (2020). *RUEG corpus*. Zenodo. https://doi.org/10.5281/zenodo.3765218

Zipser, F., & Romary, L. (2010, May). A model oriented approach to the mapping of annotation formats using standards. *Workshop on Language Resource and Language Technology Standards, LREC 2010*. https://hal.inria.fr/inria-00527799