CorpusCompass: A Tool for Data Extraction and Dataset Generation in Corpus Linguistics

Muhadj Adnan[1], Nicolo' Brandizzi[2]

[1]Arabic Linguistics, University of Bayreuth, Bayreuth, 95447, Germany, DE
[2]Department of Computer, Automation and Management Engineering, Sapienza University of Rome, 00185, Italy, IT

Abstract
As the need for effective tools in Corpus Linguistics continues to grow, particularly for under-resourced languages and nonstandard annotation tasks, specialized software has become essential for processing and analyzing large and complex datasets. This paper introduces *CorpusCompass*, a new open source tool for data extraction and dataset creation, which offers a number of functionalities for researchers interested in analyzing corpora. The tool can derive structured datasets from text annotated with custom annotation schemes, while also checking for errors and consistency. By defining custom variables of interest and annotation rules, researchers can tailor the tool to their specific needs, making it particularly valuable for unique linguistic research domains. When used in conjunction with statistical analysis or visualization tools, *CorpusCompass* helps researchers to gain insights into the factors that are affecting language use. In this paper, we introduce the tool and give a real-world example in the field of language variation.

Keywords
Corpus Linguistics, Under-resourced languages, Nonstandard annotation tasks, Data Extraction, Dataset generation, Statistical analysis, Language variation, Sociolinguistics

1. Introduction

In recent years, there has been a growing interest in studying language variation in under-resourced languages. Mair [1] identifies a lack of resources for spoken data in corpus linguistics and emphasizes the need for more computational tools for different languages and varieties. The process of creating and analyzing a spoken language corpus is complex, posing a range of challenges for researchers in the field of Corpus Linguistics.

In this context, we identify six main steps for this process, each presenting its own set of practical and technical challenges, see Figure 1. Step (i) entails *sourcing and recording data*, along with the associated metadata, which provides essential contextual information about the recordings. Following this, step (ii) involves *transcribing* the spoken data to convert it into a text-based format, allowing for more straightforward analysis. The third step (iii) involves *annotating* the data with linguistic features, such as phonological or morphological information. Subsequently, step (iv) requires *data preprocessing* to clean and organize the data, preparing it for the fifth step, (v), which is *data analysis*. This stage allows researchers to derive insights from the corpus by examining patterns and connections within the data. Lastly, the final step, (vi), involves *publishing and sharing* the corpus with the wider research community, promoting collaboration and further research based on the spoken language data.

Annotating can be a time-consuming and error-prone process, especially when working with large corpora. Errors in a manually annotated corpus can potentially affect the evaluation. Additionally, poor quality annotation in the corpus can lead to misleading results in a linguistic analysis. This is where *CorpusCompass* comes into play.

In this paper, we provide a detailed overview of *CorpusCompass*[1], including its design, implementation, and functionalities. The tool is based on *Jupyter Notebook* and is designed to help Corpus Linguistics researchers focusing on language variation to create a structured dataset from their previously annotated corpus/corpora and a list of variables of interest (see Section 3.1). The tool is coded in Python and can be run in an interactive manner using *Google Colab*. Once run, it generates a structured dataset, i.e. a systematically organized collection of data, that includes linguistic variables and potentially relevant metadata (see Section 3.2).

On the one hand, the dataset enables corpus exploration, assisting in discovering patterns that can inform the creation of new hypotheses or the dismissal of initial assumptions (see Al-Wer et al. [2], pp. 37-38). On the other hand, it facilitates performing statistical analyses using established methods through platforms like Rbrul

[1]The code is available on GitHub, for the URL, please visit the website https://www.corpuscompass.com/. Please note that the code for *CorpusCompass* is constantly evolving and, in this paper, we refer to version 1.0.0.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
muhadj.adnan@uni-bayreuth.de (M. Adnan); brandizzi@diag.uniroma1.it (N. Brandizzi)
https://nicofirst1.github.io/ (N. Brandizzi)
0009-0001-5174-3897 (M. Adnan); 0000-0002-3191-6623 (N. Brandizzi)

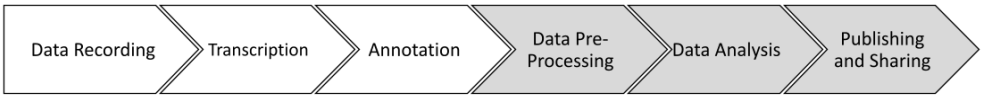


Figure 1: Practical and technical challenges when creating spoken language corpora in six steps: from sourcing and recording data (and metadata), to transcribing, annotating, and marking up datasets. The last three steps (data preprocessing, data analysis, and publishing and sharing) are highlighted, as these are the areas where *CorpusCompass* provides support and assistance to researchers.

[3], SPSS [4], and R [5].

To exemplify the practical application of *CorpusCompass* within a research context, Section 4 explores a case study on inter-generational linguistic variation among Iraqi speakers living in Germany, and highlights various potential uses of the tool.

For linguists without any or limited programming skills, *CorpusCompass* has the potential of saving time (by automating repetitive tasks) and improving the accuracy of their research. It is intended to bridge the gap between researchers and advanced statistical analysis by facilitating the connection between them, as well as addressing research questions that require the use of manually annotated data. Moreover, *CorpusCompass* was developed by researchers in Linguistics and Computer Science to advance Corpus Linguistics tools and promote interdisciplinary collaboration.

2. Related Work

Many text annotation tools have been developed primarily in the context of Artificial Intelligence and Natural Language Processing. However, such tools typically do not address specific needs, focusing more on the annotation of informational content rather than linguistic properties of text (phonological, grammatical, lexical, etc.). For this reason, linguists typically work with tools that have been developed specifically for the purpose of annotating linguistic corpora. In the following, we will briefly outline the importance of these tools and demonstrate how *CorpusCompass* complements and extends their functionality.

There are several tools available for researchers in Corpus Linguistics (see Neves and Ševa [6], Berberich and Kleiber [7]). *AntConc* [8], *Monoconc Pro* [9], and *WordSmith* [10] remain popular tools due to their wide range of functions, including KWIC (key word in context) concordancers, collocates, word frequencies, keywords and other corpus analysis features.

AntConc is a powerful corpus analysis tool, but it has certain drawbacks. One major limitation is the lack of a feature to create structured datasets from the corpus or various corpora, which is essential for statistical analysis pipelines as well as to share data. This is particularly problematic for researchers working with complex data involving various phonological, morphological and lexical variables as well as different speakers, as detailed and structured extraction is necessary. Sociolinguistic studies, in particular, require flexibility in handling multiple speakers and their background information. With *CorpusCompass*, we aim to address this gap in functionality.

WordSmith and *Monoconc Pro* share these limitations with *AntConc*, but they also have an additional drawback: they are not freely available. This lack of open access can be a significant barrier for corpus linguistics practitioners who may not have the financial resources to purchase these tools.

Other useful modern tools are typically focused on the annotation process rather than the analysis. For example, INCEpTION[2] [12] is a cloud-based platform that enables researchers to create and share linguistic annotations in a collaborative environment, for various languages. Another example is FLAT [13], a web-based linguistic annotation tool that revolves around the FoLiA format, a customizable XML-based format for linguistic annotation. However, for state-of-the-art analysis and error-checking of annotations produced by such tools, we would prefer to not rely on their built-in capabilities, but instead make use of common data science methodology such as statistical analysis in *R* or visualization with *Gephi*. This requires exporting the annotations from the typically XML-like formats that these tools use to a tabular data format used in data science such as CSV or TSV files. This is a core functionality of *CorpusCompass*, addressing a gap in the existing tools.

Biber et al. [14], Gries [15], and Weisser [16] suggest that learning programming and developing one's own analytical tools can overcome limitations in existing corpus tools.

Biber et al. [14] suggest that this would allow corpus linguists to perform faster and more accurate analyses, and the ability to tailor the output to suit the particular research requirements. Furthermore, according to Gries [15], utilising a pre-existing tool lets the researcher become dependent of the company or individual developing them, whereas programming allows them to have control over their research needs. Therefore, corpus lin-

[2]The software was previously known as *WebAnno* [11].

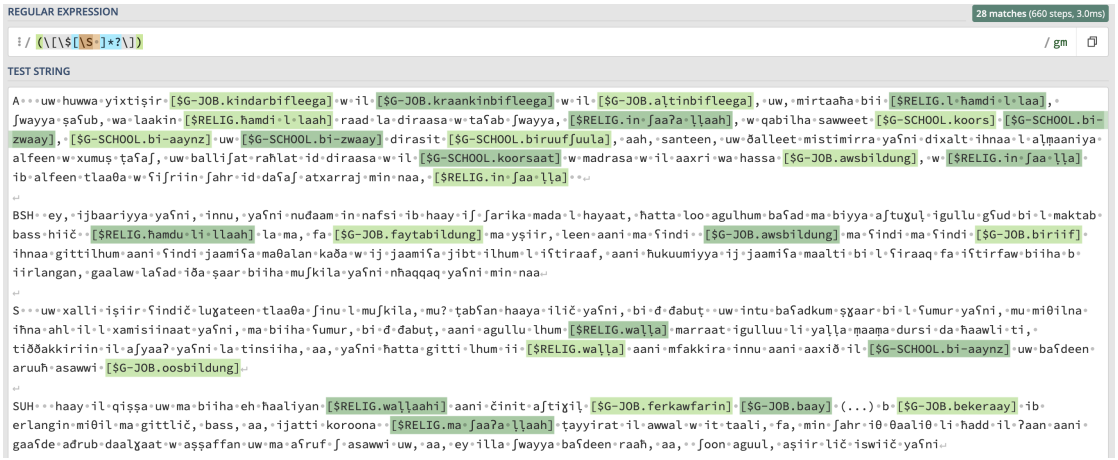


Figure 2: Example of a subset of the corpus dealing with linguistic variation, highlighting the annotations matching the standard *regex* pattern in green. Note the different anonymized speakers based on age: young (A,BSH); old (S, SUH). Image made with https://regex101.com/.

guists have clear benefits from learning a programming language, both in terms of the flexibility to develop tools for specialized tasks, as well as providing them with an understanding of the issues faced by tool developers creating general purpose tools.

As seen in the overview above, each tool addresses unique needs, thus implementing specific functionalities. In comparison, *CorpusCompass* tackles a complementary set of challenges. It is implemented as a *Jupyter Notebook* and can be run in an interactive manner using *Google Colab*, which not only makes it more accessible to users but also gives beginners the possibility to get familiar with programming.

3. Methodology

The modular code structure of *CorpusCompass* includes a file handling module for managing JSON and CSV files, an annotation parsing module that extracts annotations using regular expressions (refer to Section 3.1), a dataset construction module for creating a structured CSV dataset, and a logging module that displays relevant information such as program status and annotation details. *CorpusCompass* provides helpful functions for string manipulation, data cleaning, and error handling, simplifying the data analysis process. These functions enhance dataset accuracy by mitigating errors and inconsistencies during the data preparation phase. In the following, we describe the pipeline of *CorpusCompass* and demonstrate how it simplifies the process of extracting valuable insights from spoken language data.

3.1. Defining Variables

Defining variables of interest is a crucial step in using *CorpusCompass*, as it allows researchers to tailor the tool to what they aim to investigate. In the field of Corpus Linguistics, variables are often used to study language variation and how it is affected by various factors such as speaker demographics (e.g. age, sex, education), linguistic context (e.g. dialect, register), social context (e.g. audience, situation), or properties of a construction (e.g. morphemes, idioms). By defining both independent and dependent variables in their structured dataset, researchers have maximum freedom in the exploration of their corpora and the creation of their unique datasets.

Regular Expressions Based on their research objectives, researchers may use automated annotation tools or choose to manually annotate data in more complex linguistic situations (as described in Section 4). This leads to a broad range of annotation rules. To accommodate this variety, *CorpusCompass* employs regular expressions (*regex*) [17] for accurate extraction of annotations from the corpus. *Regex* allows the user to define text patterns, and is useful for tasks such as input validation, text modifications, and data extraction.

Figure 2 shows four paragraphs taken from our corpus[3], where annotations are highlighted in green. For the sake of simplicity, we kept only the dependent variable that are at the basis of our analysis in Section 4.2.

[3]Strict phonetic transcription was not followed due to the focus on pre-selected specific features, as adhering to it for the extensive 22-hour audio recordings would have been time-consuming.

Appendix A reports the full list of variables used for the study. It is important to note that in our complete annotated files, we typically have multiple annotations per word.

3.2. Generating a Structured Dataset

After the variables and *regex* rules have been defined, researchers can run *CorpusCompass* to generate a structured dataset. The dataset construction module automatically performs several steps, including cleaning and preprocessing the extracted annotations, grouping them by speaker and file, and writing them to a CSV file. Six output files are created, including five CSVs (dataset, annotation_info, missed_annotations, unk_variables, and binary_dataset) and one JSON file (corpus_stats).

The *dataset* file is a structured dataset based on the defined variables, with each row corresponding to a token in the annotated corpus and each column representing a variable category. This organized representation of annotations facilitates the analysis. The *annotation_info* file contains information about the annotations included in the dataset file, such as the token itself, the number of times it appears in the dataset, and the number of times it appears for each speaker. This information can be useful for identifying patterns in the data, such as the most common tokens or the distribution of annotations across speakers.

The *missed_annotations* file tracks tokens that were previously annotated but not consistently annotated in subsequent instances[4]. The file's purpose is to identify tokens that were once deemed important but not annotated consistently. Furthermore, in projects with multiple annotators, inter-annotator disagreement is a known challenge [18]. During the annotation process, it is possible for researchers to come up with new variables that were not previously specified in the JSON. However, researchers may forget to add these variables to the file, leading to inconsistencies in the dataset. The *unk_variables* file contains a list of variables that were not specified in the JSON. Finally, the *corpus_stats* file provides an overview of the corpus by reporting key statistics. Access to these statistics enables researchers to better understand the size and structure of their corpus, and can also provide valuable information for reproducibility purposes.

For a comprehensive description and additional information regarding the CSV files, please see Appendix B.

4. Analysing Linguistic Variation Using *CorpusCompass*

Following the annotation of the data with linguistic features, we used *CorpusCompass* for preprocessing in order to prepare the data for the analysis process. This transition from data preprocessing to analysis was facilitated by the integration of the tool into our workflow, significantly enhancing the efficiency and effectiveness of our exploration process.

In our case study, we examine Arabic-speaking communities, specifically Iraqis and Syrians, residing in Germany since 2014, following standard sociolinguistic variationist research practices. The participants are Iraqi and Syrian Arabic native speakers. We select phonological, morphological, and lexical variables for statistical analysis, with age as a key independent variable influencing linguistic variation. The study aims to examine inter-generational differences within the two groups and explore the extent to which a koiné (common variety) results from dialect and language contact in the migration context between the Syrian and Iraqi participants.

4.1. The Corpus

The corpus utilized in this study, of which a sample is illustrated in Figure 2, represents only half of our entire dataset and has been phonetically transcribed[5] using the International Phonetic Alphabet (IPA) and annotated by a single person in *Notepad++*. While having one annotator can be a common case in the field, mostly due to resource limitation, it can also be prone to errors and inconsistencies. The analyzed corpus is comprised of 2,101 paragraphs and encompasses 114,550 words and 654,431 characters. It features 24 speakers in total, 14 of whom are Iraqis, considered speakers of interest for our analysis. The dataset contains 35 variables, with 25 being dependent variables and 10 being independent variables. These variables are represented by 69 distinct values, 53 of which correspond to dependent variable values, and 16 to independent variable values.

In total, the corpus contains 3,366 unique annotations, with 13,641 annotated tokens. Given the substantial size of the dataset and the numerous variables involved, organizing the data in a structured manner is crucial for efficient analysis.

For comprehensive details regarding the corpus collection process, transcription methodology, annotation procedures, and the specific tool employed, please refer to Appendix C.

[4]The file might contain false positives since *CorpusCompass* does not differentiate between different meanings of the same token.

[5]Supported by one assistant during transcription of the recorded data.

Research Question In the following section, we use *CorpusCompass* to answer two research questions focusing on Iraqi speakers: (i) does age influence the usage of religious expressions? (ii) are young speakers more subject to German borrowings while speaking Arabic? These questions will guide our exploration of potential correlations within the dataset, with the understanding that the current analysis serves as a simplified demonstration. However, it should be noted that the dataset is well-suited for rigorous statistical analysis, including techniques such as regression analysis.

Dependent Variables For the purpose of this example, we have selected two categories of dependent variables to investigate: (i) religious expressions[6], represented by the label RELIG; (ii) the influence of German language, represented by multiple labels such as G-DL for daily life, G-EDU for education, and G-JOB for working contexts. Since we are interested in the general use of German words, we generalize the labels to GERM.

Independent Variable The independent variable chosen for analysis in the corpus is *age*, which is an important factor to consider in language variation. The speakers are divided into two categories, young (21-26 years) and old (46-55 years).

4.2. Analyzing and Sharing the Data

This section discusses two types of analysis: error checking and data analysis. Error checking is the process of identifying and fixing errors or inconsistencies in the data, while data analysis involves using statistical and visualization techniques to extract insights and draw conclusions from the data.

Error Checking *CorpusCompass* can identify any errors or inconsistencies during the annotation process and generate separate CSV files that provide information on annotated and non-annotated tokens. Thanks to the generated file, we were able to find circa 400 (3% of all the annotations) ill-formatted annotations (e.g. "[$G-JOB.biriif["), 1,305 missed annotations of which 205 where considered correctly identified and more than 9 unknown variables, i.e. dependent variables that are present in the corpus but not specified beforehand by the user.

Data Analysis By importing *the binary_dataset* in Excel [21], we determined the cross-tabulation (pair-wise frequency) of dependent and independent variables. Table 1 presents these frequencies along with the total number of words spoken by young and old speakers. By normalizing the frequencies and estimating the proportions (old vs. young), we observe the following:

Table 1
Frequency of dependent variables (*GERM* and *RELIG*) across age groups (Old and Young) along with the total number of words spoken by each group.

Age Group	GERM	RELIG	Words
Old	221	357	42,483
Young	505	175	39,406

$$RELIG_p = \frac{RELIG_{old}}{RELIG_{yng}} \cdot \frac{Words_{yng}}{Words_{old}} = 1.89$$

This indicates that old speakers use 189% times more religious phrases than young speakers. Applying the same method for GERM, we obtain:

$$GERM_p = \frac{GERM_{old}}{GERM_{yng}} \cdot \frac{Words_{yng}}{Words_{old}} = 0.40$$

Correspondingly, old speakers use 40% of the amount of German borrowings compared to young speakers. To assess the significance of our findings, we conducted standard t-test analyses with DATAtab [22]. When comparing the proportion of older speakers using religious expressions to that of younger speakers using the same expressions, the result was the following:

$$t(10695) = -7.91,\ p < .001$$

In contrast, the analysis of German borrowings between older and younger populations yielded:

$$t(10695) = 10.59,\ p < .001$$

The *p-value* suggests that the dependent variables (German borrowings and religious expressions) play a role in the language variation exhibited by young and old Iraqi migrants residing in Germany and requires further investigation. Ultimately, these findings validate our initial research question and demonstrate the value of structured data in facilitating robust statistical analyses.

5. Limitations

CorpusCompass, while offering numerous features, is not without its limitations. In this section, we outline some of the primary constraints of the tool, alongside potential future developments.

[6] Jaradat [19] and Piamenta [20] describe religious phrases, such as *Inshallaah* (God willing), *alhamdulillaah* (Praise be to God), *Allah ysallimak* (may God protect you) etc. as "Allah expressions". They include an explicit or implicit reference to Allah, which is literally translated as "the God".

The user interface, especially the integration of *regex* and *Jupyter Notebook* might pose a challenge to linguists, particularly those hesitant to engage with programming. This could be solved by developing the tool into a more user-friendly application. Furthermore, we showcased the functionality of *CorpusCompass* in addressing a specific research question. While the tool has already been applied to address other research questions [23] and on another corpus[7], the extend of its usability remains a point of investigation. It is essential to assess its performance on multiple corpora to enhance its robustness and confirm its applicability for diverse research contexts.

Additionally, while the tool identifies errors, as delineated in Section 4.2, the manual correction process can be tedious and time-consuming. Looking forward, future iterations of *CorpusCompass* might integrate an automatic error correction feature that suggests possible corrections, allowing users to either accept or decline them. Another area for consideration is the tool's reliance on the CSV format, which might present compatibility issues with other linguistic tools. Transitioning to more standardized formats, such as XML, upcoming versions could address this limitation.

In summary, there are numerous opportunities to refine and enhance *CorpusCompass*. By addressing its current constraints, introducing new functionalities, and emphasizing user-centric enhancements, this tool has the potential to become an even more invaluable asset in Corpus Linguistics.

6. Conclusion

Creating and analyzing a corpus is a complex task that requires a range of technical and practical skills. In this paper, we have explored the challenges involved in these steps and introduced *CorpusCompass* as an innovative solution. The tool's aim is to simplify data extraction and dataset generation, facilitating the identification of significant features and syntactic errors in the annotations. This contributes to advancing the overall replicability of studies within the field of Corpus Linguistics. As *CorpusCompass* is implemented as a *Jupyter Notebook*, it also serves as an accessible introduction to programming for researchers who wish to expand their skill set and gain more control over their analytical processes.

Additionally, we have presented a real-world example of how *CorpusCompass* can be applied in the field of language variation by using a subset of our corpus of Arabic varieties spoken by migrants in Germany, representing an under-resourced language. The example shows how the generated dataset can be used in conjunction with existing analysis tools to answer unique research questions. With *CorpusCompass*, we aim to contribute to the development of tools for spoken language corpora. The existence of this tool and its accessibility to researchers without a background in programming will lead to more quantitative studies that analyse such corpora. The tool exemplifies interdisciplinary collaboration and emphasizes the importance of linguistics researchers working with experts from computer science and engineering. This collaboration results in the development of flexible corpus tools applicable to a wide range of research studies.

Sharing is Caring It is essential to highlight that structured datasets are crucial for sharing data in linguistic research, particularly in connection to research data management practices and platforms such as *Figshare* [24]. Organized data facilitates sharing and reusability among researchers, enabling more extensive collaborations and the creation of larger datasets. Furthermore, structured datasets allow researchers to replicate and verify research findings, promoting transparency and accountability in the scientific community. Therefore, creating a structured dataset is not only essential for internal analysis but also for the advancement of the field and the dissemination of knowledge.

[7] The corpus is focused on Nigerian Arabic and has been kindly provided by Prof. Dr. Jonathan Owens.

Acknowledgments

We extend our sincere appreciation to Dr. Jelke Bloem (University of Amsterdam) and Prof. Dr. Jonathan Owens (University of Bayreuth) for their invaluable feedback.

The development of *CorpusCompass* was undertaken as part of the project *Modernity, Migration, and Minorities: Three Case Studies of Arabic in Contact* at the University of Bayreuth and funded by the Deutsche Forschungsgemeinschaft (DFG), No. 429257272.

References

[1] C. Mair, Erfolgsgeschichte korpuslinguistik?, in: Korpuslinguistik, De Gruyter, 2018, p. 5–26. URL: http://dx.doi.org/10.1515/9783110538649-002. doi:10.1515/9783110538649-002.

[2] E. Al-Wer, U. Horesh, B. Herin, R. De Jong, Arabic Sociolinguistics, Cambridge University Press, 2022. doi:10.1017/9781316863060.

[3] D. E. Johnson, Getting off the GoldVarb standard: Introducing rbrul for mixed-effects variable rule analysis, Language and Linguistics Compass 3 (2009) 359–383. URL: https://doi.org/10.1111/j.1749-818x.2008.00108.x. doi:10.1111/j.1749-818x.2008.00108.x.

[4] H. Norman, C. H. Hull, SPSS: Statistical package for the social sciences, McGraw-Hill Book Company, 1975.

[5] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. URL: https://www.R-project.org/.

[6] M. Neves, J. Ševa, An extensive review of tools for manual annotation of documents, Briefings in Bioinformatics 22 (2019) 146–163. URL: https://doi.org/10.1093/bib/bbz130. doi:10.1093/bib/bbz130.

[7] K. Berberich, I. Kleiber, Tools for corpus linguistics, https://corpus-analysis.com/ (Mar. 2023), 2020.

[8] L. Anthony, Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit, proceedings of IWLeL (2004) 7–13.

[9] A. Svedkauskaite, Monoconc pro 2.0 and the corpus of spoken professional american english: Resources from athelstan, Style 38 (2004) 127–133.

[10] S. Mike, Wordsmith tools version 8, 2023.

[11] R. E. De Castilho, E. Mújdricza-Maydt, S. M. Yimam, S. Hartmann, I. Gurevych, A. Frank, C. Biemann, A web-based tool for the integrated annotation of semantic and syntactic structures, in: Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH), 2016, pp. 76–84.

[12] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The inception platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/, event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

[13] M. van Gompel, M. Reynaert, Folia: A practical xml format for linguistic annotation–a descriptive and comparative study, Computational Linguistics in the Netherlands Journal 3 (2013) 63–81.

[14] D. Biber, S. Conrad, R. Reppen, Corpus linguistics: Investigating language structure and use, Cambridge University Press, Shaftesbury Road ,Cambridge, CB2 8BS , United Kingdom, 1998.

[15] S. T. Gries, What is corpus linguistics?, Language and linguistics compass 3 (2009) 1225–1241.

[16] M. Weisser, Essential programming for linguistics, Edinburgh University Press, The Tun - Holyrood Road. 12 (2f) Jackson's Entry. Edinburgh EH8 8PJ. UK., 2009.

[17] A. V. Aho, Algorithms for finding patterns in strings, in: J. VAN LEEUWEN (Ed.), Algorithms and Complexity, Handbook of Theoretical Computer Science, Elsevier, Amsterdam, 1990, pp. 255–300. URL: https://www.sciencedirect.com/science/article/pii/B9780444880710500102. doi:https://doi.org/10.1016/B978-0-444-88071-0.50010-2.

[18] Y. Oortwijn, T. Ossenkoppele, A. Betti, Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks, in: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Online, 2021, pp. 131–141. URL: https://aclanthology.org/2021.humeval-1.15.

[19] A. A. Jaradat, The linguistic variants of allah expressions in jordanian arabic, Cross-Cultural Communication 10 (2014) 61–68.

[20] M. Piamenta, The Muslim conception of God and human welfare: As reflected in everyday Arabic speech, Brill Archive, Leiden, The Netherlands, 1983. URL: https://brill.com/view/title/1464. doi:https://doi.org/10.1163/9789004661820.

[21] C. Microsoft, Microsoft excel, https://office.microsoft.com/excel (Apr. 2023), 2023. URL: https://office.microsoft.com/excel.

[22] T. DATAtab, Datatab: Online statistics calculator, https://datatab.net/ (Apr. 2023), 2023. URL: https://datatab.net/.

[23] M. Adnan, J. Owens, Imperfect Verbal Prefixes as Discourse Markers, volume XXXV, Perspectives on Arabic Linguistic, Benjamins, Amsterdam, To Appear.

[24] T. Figshare, Figshare: the open research repository platform, 2023. URL: https://figshare.com/.

[25] L. Milroy, M. Gordon, Sociolinguistics: Method and interpretation, Wiley, 2003. URL: https://doi.org/10.1002/9780470758359. doi:10.1002/9780470758359.

[26] S. A. Tagliamonte, Analysing sociolinguistic variation, Cambridge University Press, 2006.

[27] P. Boersma, D. Weenink, Praat: Doing phonetics by computer, Ear and Hearing 32 (2011). URL: https://journals.lww.com/ear-hearing/Fulltext/2011/03000/Praat__Doing_Phonetics_by_Computer.12.aspx.

Speakers

```
{
  "A" : ["female","young"],
  "BSH" : ["male","young"],
  "S" : ["female","old"],
  "SUH" : ["female", "old"]
}
```

Independent variables

```
{
  "Gender":["male","female"],
  "Age":["old","young"]
}
```

Dependent variables

```
{
  "German Context" : [ "G-SCHOOL",  "G-JOB"],
  "Religious Phrases" : "RELIG"
}
```

Figure 3: Content of variables and speaker JSON files.

A. Variables

The study presented in Section 4.1 revolves around the variables and speaker information detailed in Table 3. The employed syntax adheres to the JSON specification, providing considerable flexibility in the examination of linguistic variables.

There are two main types of variables: independent and dependent variables. Independent variables, also known as input variables, are the factors that the researcher manipulates or controls in a study. In contrast, dependent variables, also known as output variables, are the outcomes or responses being measured. The dependent variables are affected by the independent variables.

The speakers involved in the example are represented by four anonymized aliases. For each individual, two attributes (i.e. independent variables), namely *age* and *gender*, are taken into consideration. These attributes encompass the categories of *male* and *female*, as well as *old* and *young*. Finally, the corpus is annotated with two dependent variables: *german context*, which can be categorized as either G-SCHOOL or G-JOB, and *religious phrases*, identified by the annotation RELIG.

B. Generated Files

Additional details regarding the files generated by *CorpusCompass* are provided in the following. In order to maintain simplicity and avoid overwhelming the reader with excessive information, we conducted calculations on a subset corpus, as illustrated in Figure 2. This facilitates a visual association between the output of *CorpusCompass* and the input data. Conversely, the files containing missed annotations and unknown variables are presented for the entire corpus, which serves as the foundation for the analysis in Section 4.2. This distinction is necessary as the subset corpus exhibits no errors or missed annotations.

Dataset The *dataset* file encompasses a structured dataset that reflects the defined variables, where each row corresponds to a token within the annotated corpus, and each column represents a distinct variable category (see Figure 4). Additionally, a *binary_dataset* file contains a one-hot encoded version of the dataset, specifically designed for seamless integration with statistical models and machine learning pipelines, without necessitating additional preprocessing steps.

Within the provided CSV example, Figure 4, *token* refers to the individual tokens in the corpus, *German Context* indicates the context of the text (either G-JOB or G-SCHOOL), *Religious Phrases* denotes the presence of religious phrases (annotated with RELIG), *age* and *gender* specify the respective attributes of the speaker, *speaker* identifies the speaker's anonymized alias, *interlocutor/s* denotes the interlocutor(s) in the conversation, *file* points to the file path where the token was spotted (truncated in the example), and *context* provides additional contextual information.

For instance, the token *kindarbifleega* has G-JOB as the German context, no religious phrases, a young female speaker (age and gender), identified as *A*, and an anonymized speaker-interlocutor combination of *BSH*, *S*, *SUH*. The corresponding context is *A uw huwwa yixtiṣir kindarbifleega...* .

token	German Context	Religious Phrases	Age	Gender	speaker	interlocutor/s	file	context
kindarbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	A uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaha bii
kraankinbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	A uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaha bii [$RELIG.ḥ hamdi l
aḷtinbifleega	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	uw huwwa yixtiṣir kindarbifleega w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaha bii ḥ hamdi l laa , ʃwayya ṣaʕub,
ḥ hamdi l laa		RELIG	young	female	A	BSH,S,SUH	Path/2/file	w il kraankinbifleega w il aḷtinbifleega , uw, mirtaaha bii ḥ hamdi l laa , ʃwayya ṣaʕub, wa laakin [$RELIG.ḥamdi l
ḥamdi l laah		RELIG	young	female	A	BSH,S,SUH	Path/2/file	bii ḥ hamdi l laa , ʃwayya ṣaʕub, wa laakin ḥamdi l laah raad la diraasa w taʕab ʃwayya, [$RELIG.in
in ʃaa?a ‖aah		RELIG	young	female	A	BSH,S,SUH	Path/2/file	ḥamdi l laah raad la diraasa w taʕab ʃwayya, in ʃaa?a ‖aah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw
koors	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	w taʕab ʃwayya, in ʃaa?a ‖aah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw
bi-zwaay	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	taʕab ʃwayya, in ʃaa?a ‖aah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ð
bi-aaynz	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	ʃwayya, in ʃaa?a ‖aah , w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet r
bi-zwaay	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	ʃaa?a ‖aah], w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet mistimirra y
biruufʃuula	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	w qabilha sawweet koors bi-zwaay , bi-aaynz uw bi-zwaay dirasit biruufʃuula , aah, santeen, uw ðalleet mistimirra yaʕni dixalt ih
koorsaat	G-SCHOOL		young	female	A	BSH,S,SUH	Path/2/file	xumuṣ ṭaʕaʃ, uw balliʃat rahlat id diraasa w il koorsaat w madrasa w il aaxri wa hassa awsbildung , w
awsbildung	G-JOB		young	female	A	BSH,S,SUH	Path/2/file	il koorsaat w madrasa w il aaxri wa hassa awsbildung , w in ʃaa ‖a ib alfeen tlaaθa w ʕiʃriin
in ʃaa ‖a		RELIG	young	female	A	BSH,S,SUH	Path/2/file	w madrasa w il aaxri wa hassa awsbildung , w in ʃaa ‖a ib alfeen tlaaθa w ʕiʃriin ʃahr id
in ʃaa ‖a		RELIG	young	female	A	BSH,S,SUH	Path/2/file	tlaaθa w ʕiʃriin ʃahr id daʕaʃ atxarraj min naa, in ʃaa ‖a
ḥamdu li llaah		RELIG	young	male	BSH	A,S,SUH	Path/2/file	biyya aʃtuχuḷ igullu gʕud bi l maktab bass hiič ḥamdu li llaah la ma, fa faytabildung ma yṣiir, leen
faytabildung	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	maktab bass hiič ḥamdu li llaah la ma, fa faytabildung ma yṣiir, leen aani ma ʕindi awsbildung ma ʕindi
awsbildung	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	ma, fa faytabildung ma yṣiir, leen aani ma ʕindi awsbildung ma ʕindi ma ʕindi biriif ihnaa gittiihum aani ʕindi
biriif	G-JOB		young	male	BSH	A,S,SUH	Path/2/file	leen aani ma ʕindi awsbildung ma ʕindi ma ʕindi biriif ihnaa gittiihum aani ʕindi jaamiʕa maθalan kaða w ij
waḷ‖a		RELIG	old	female	S	A,BSH,SUH	Path/2/file	ma biiha ʕumur, bi ɗ ɗabuṭ, aani agullu lhum waḷ‖a marraat igulluu li yaḷ‖a ṃaaṃa dursi da ḥaawli ti,
waḷ‖a		RELIG	old	female	S	A,BSH,SUH	Path/2/file	yaʕni la tinsiiha, aa, yaʕni ḥatta gitti lhum ii waḷ‖a aani mfakkira innu aani aaxið il bi-aaynz uw baʕdeen
bi-aaynz	G-SCHOOL		old	female	S	A,BSH,SUH	Path/2/file	lhum ii waḷ‖a aani mfakkira innu aani aaxið il bi-aaynz uw baʕdeen aruuh asawwi oosbildung
oosbildung	G-JOB		old	female	S	A,BSH,SUH	Path/2/file	innu aani aaxið il bi-aaynz uw baʕdeen aruuh asawwi oosbildung
waḷ‖aahi		RELIG	old	female	SUH	A,BSH,S	Path/2/file	SUH haay il qiṣṣa uw ma biiha eh ḥaaliyan waḷ‖aahi aani činit aʃtiɣil ferkawfarin baay (...) b bekeraay ib
ferkawfarin	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	uw ma biiha eh ḥaaliyan waḷ‖aahi aani činit aʃtiɣil ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič,
baay	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	ma biiha eh ḥaaliyan waḷ‖aahi aani činit aʃtiɣil ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič, bass,
bekeraay	G-JOB		old	female	SUH	A,BSH,S	Path/2/file	ḥaaliyan waḷ‖aahi aani činit aʃtiɣil ferkawfarin baay (...) b bekeraay ib erlangin miθil ma gittlič, bass, aa, ijatti koroona
ma ʃaa?a ‖aah		RELIG	old	female	SUH	A,BSH,S	Path/2/file	ib erlangin miθil ma gittlič, bass, aa, ijatti koroona ma ʃaa?a ‖aah ṭayyirat il awwal w it taali, fa,

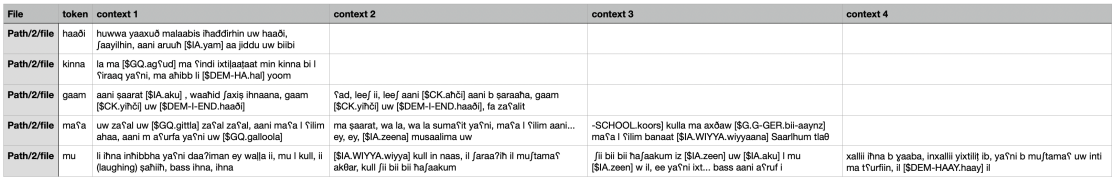
Figure 4: Dataset file generated from example corpus in Figure 2.

token	annotated	not annotated	not_annotated_interest	A not annotated	A annotated	BSH not annotated	BSH annotated	S not annotated	S annotated	SUH not annotated	SUH annotated	total
kindarbifleega	1	0	0	0	1	0	0	0	0	0	0	1
kraankinbifleega	1	0	0	0	1	0	0	0	0	0	0	1
aḷtinbifleega	1	0	0	0	1	0	0	0	0	0	0	1
ḥ hamdi l laa	1	0	0	0	1	0	0	0	0	0	0	1
ḥamdi l laah	1	0	0	0	1	0	0	0	0	0	0	1
in ʃaa?a ‖aah	1	0	0	0	1	0	0	0	0	0	0	1
koors	1	0	0	0	1	0	0	0	0	0	0	1
bi-zwaay	2	0	0	0	2	0	0	0	0	0	0	2
bi-aaynz	2	0	0	0	1	0	0	0	1	0	0	2
biruufʃuula	1	0	0	0	1	0	0	0	0	0	0	1
koorsaat	1	0	0	0	1	0	0	0	0	0	0	1
awsbildung	2	0	0	0	1	0	1	0	0	0	0	2
in ʃaa ‖a	2	0	0	0	2	0	0	0	0	0	0	2
ḥamdu li llaah	1	0	0	0	0	0	1	0	0	0	0	1
faytabildung	1	0	0	0	0	0	1	0	0	0	0	1
biriif	1	0	0	0	0	0	1	0	0	0	0	1
waḷ‖a	2	0	0	0	0	0	0	0	2	0	0	2
oosbildung	1	0	0	0	0	0	0	0	0	1	0	1
waḷ‖aahi	1	0	0	0	0	0	0	0	0	0	1	1
ferkawfarin	1	0	0	0	0	0	0	0	0	0	1	1
baay	1	0	0	0	0	0	0	0	0	0	1	1
bekeraay	1	0	0	0	0	0	0	0	0	0	1	1
ma ʃaa?a ‖aah	1	0	0	0	0	0	0	0	0	0	1	1

Figure 5: Annotation information file generated from example corpus in Figure 2.

Annotation Information The *annotation_info* file contains information about the annotations included in the *dataset* file, such as the token itself, the number of times it appears in the dataset, and the number of times it appears for each speaker, Figure 5.

Missed Annotations The *missed_annotations* file tracks tokens that were previously annotated but not consistently annotated in subsequent instances. It contains the token and its context, determined by a user-defined n-gram size. Figure 6 reports an example taken from the from corpus in Section 4.1.



File	token	context 1	context 2	context 3	context 4
Path/2/file	haaði	huwwa yaaxuð malaabis ihaðdirhin uw haaði, ʃaayilhin, aani aruuh [$IA.yam] aa jiddu uw biibi			
Path/2/file	kinna	la ma [$GQ.agʔud] ma ʕindi ixtiʃaaṭaat min kinna bi l ʕiraaq yaʔni, ma ahibb li [$DEM-HA.hal] yoom			
Path/2/file	gaam	aani ṣaarat [$IA.aku] , waaḥid ʃaxiṣ ihnaana, gaam [$CK.yiħčI] uw [$DEM-I-END.haaði]	ʔad, leeʃ ii, leeʃ aani [$CK.aħčI] aani b ṣaraaḥa, gaam [$CK.yiħčI] uw [$DEM-I-END.haaði], fa zaʔalit		
Path/2/file	maʔa	uw zaʔal uw [$GQ.gittla] zaʔal zaʔal, aani maʔa l ʕilim ahaa, aani m aʕurfa yaʔni uw [$GQ.galloola]	ma ṣaarat, wa la, wa la sumaʔit yaʔni, maʔa l ʕilim aani... ey, ey, [$IA.zeena] musaalima uw	-SCHOOL.koors] kulla ma axðaw [$G.G-GER.bii-aaynz] maʔa l ʕilim banaat [$IA.WIYYA.wiyyaana] Saarlhum tlaθ	
Path/2/file	mu	li ihna inḥibbha yaʔni daaʔiman ey waḷa ii, mu l kull, ii (laughing) ṣaḥiih, bass ihna, ihna	[$IA.WIYYA.wiyya] kull in naas, il ʃaraaʔiħ il muʃtamaʕ akθar, kull ʃii bii bii haʃaakum	ʃii bii bii haʃaakum iz [$IA.zeen] uw [$IA.aku] l mu [$IA.zeen] w il, ee yaʕni ixt... bass aani aʕruf i	xalili ihna b ɣaaba, inxallii yixtiliṭ ib, yaʕni b muʃtamaʕ uw inti ma tʕurfiin, il [$DEM-HAAY.haay] il

Figure 6: Subset of missed annotation file generated from corpus in Section 4.1.

variable	speaker	context	token	tag
KQ	SUH	l leel sawweetha laffeetha uu θaani yoom gaʕadit min wakit iṣ ṣubiħ, ðabbeetha ʕa naar axaðitha haare ilhum axaðit	wakit	[$KQ.wakit]
QK	A	hamm ib nafs il waqit innu ma ansa luɣati l ʕarabiyya, wiya , ey mumkin	waqit	[$QK.waqit]
QK	BSH	li l kull, wugaf bass ili, leen aani ʕindi waqit maḥduud,	waqit	[$QK.waqit]
QK	DUN	awakkilhum uu aɣayyirilhum, aaxuð waqit akθar waqti wiyya binti , adarrisha uu ahtamm biiha uu anayyimha	waqti	[$QK.waqti]
QK	DUN	beet hammeen ey, ikuun waqti malyaan, ey	waqti	[$QK.waqti]
RAISE	A	baʕad santeen uu atxarraj bass aani haaða l qisim ħabbeeta uu huwwa jidiid	atxarraj	[$RAISE.atxarraj]
GA	A	iħna gaaʕdiin il diktoora gaaʕda tiʃraħ	gaaʕda tiʃraħ	[$GA.gaaʕda tiʃraħ]
SUF-NO-H	A-G	hiyya il killa iṭ ṭarafeen	killa	[$SUF-NO-H.killa]
D-DH	S-G	daaʔiman ifaadaat bi ḍ diwal fa kill ma yruuħ yaaxudna maʕaa b ayy dawla fa it?aqlamna	yaaxudna	[$D-DH.OTH-DIAL.yaaxudna]

Figure 7: Subset of unknown variables file generated from corpus in Section 4.1.

Unknown Variables The *unk_variables* file is designed point to a list of variables that were not specified in the JSON file. This file includes information about the speakers, the context, and the file it was taken from, making it easy for researchers to identify and correct any inconsistencies in the dataset, Figure 7.

Descriptive Statistics Knowing basic descriptive statistics is fundamental in language research. The *corpus_stats* JSON file provides an overview of the corpus by reporting key statistics. The file contains four types of information: (i) word-related information such as the number of paragraphs[8], words, and characters; (ii) variable information, including the number of dependent and independent variables and their values; (iii) speaker-wise information, such as the total number of speakers, speakers of interest, and words spoken per speaker; and (iv) annotation-wise information, such as the number of unique annotations and annotated tokens, see Table 2.

C. Corpus

The study is based on 20 sociolinguistic individual interviews (circa 60 minutes each) conducted in Bayreuth and Nuremberg, located in Bavaria, Germany. Additionally, two group conversations were recorded with the same speakers (90 minutes per interview), where four Iraqi and Syrian speakers were paired together.

[8]In our example, each paragraph is a turn-taking component.

The individuals in each group come from the same dialect area and almost all come from the same circle of friends/family. In order to minimize possible influences on the interview conversations, the interviews were conducted by two assistants who are native speakers of the respective varieties.

Since the sociolinguistic interview is used as a basic tool in the study of sociolinguistic variation and it is the most common method for collecting sociolinguistic data [25], the research data were collected using this method. The goal was to move from general and impersonal questions to more specific and personal questions. Questions on selected topics encouraged respondents to narratively talk about their personal experiences (e.g., life in Germany/home country, refugee experience, friends/family, fears and concerns). Thus, the speaker's natural language could be elicited [26]. After data collection, phonetic transcription of the recordings was performed using the transcription program *Praat* [27]. Demographic information, as well as details about the respondents' backgrounds and environments, were also collected. In addition, questionnaires were employed to gather data on the interviewees' language contact behavior with speakers of other languages and language varieties.

Table 2
Corpus statistics file generated from example corpus in Figure 2.


```
"paragraphs": 4,
"speakers_of_interest": 4,
"all_speakers": 4,
"dependent_variables": 2,
"independent_variables": 6,
"variables": 8,
"variables_values": 18,
"dependent_variable_values": 10,
"independent_variable_values": 8,
"words": 333,
"characters": 1897,
"unique_annotations": 22,
"annotated_tokens": 26,
"speaker_num_words": {
    "A": 107,
    "SUH": 75,
    "S": 81,
    "BSH": 90
    }
```