Project Plan

Personalized AI Trainer for Learning German

194.207: Generative AI

Group: 62

Hamza Khaldoon Alsbaihi
e11928965@student.tuwien.ac.at
Dragana Naceva
e12308467@student.tuwien.ac.at
Nathanael Nussbaumer
e11777987@student.tuwien.ac.at
Ruby Antorufa Zöhrer
e11777751@student.tuwien.ac.at

November 19, 2025

Contents

1 Users

The target users of this system are individual language learners, in particularly adult learners, such as university students and working adults, who are learning German as a foreign language. Most of them already engage with German through university courses, online platforms, or self-study resources. Their primary goal is not only to memorize isolated vocabulary items, but to integrate new language into active use: understanding how words and structures appear in authentic sentences, and being able to apply them in conversations, writing, or exams. To support this, learners typically collect a mix of materials from their daily life, such as:

- personal vocabulary lists (in notebooks, spreadsheets, flashcards, or apps like Anki),
- notes from lectures, language courses, as well as personal conversations and chat messages
- excerpts from textbooks, online articles, or PDFs with grammar explanations,
- example sentences or phrases they encounter in movies, podcasts, or conversations.

In their current workflow, these resources are managed manually: learners add items to their lists, look up example sentences in dictionaries or on the web, and consult grammar notes when needed. This is time-consuming and often leads to fragmented practice, as vocabulary, examples, and grammar explanations are rarely brought together in a coherent way for personalized exercises.

One of the project contributors is personally learning German, so the system will be designed and evaluated using real learner data (vocabulary lists, notes, and study materials) from this ongoing learning process. This makes the system both practically grounded and closely aligned with authentic learner needs.

2 Data

Learners usually store personal vocabulary lists or course notes in notebooks, spreadsheets, flashcards, or apps such as Anki.

The description of the data could be as follows:

- **Granularity:** Mixed granularity, some notes such as vocabulary might be atomic, while others such as grammars explanation could be long documents that need to be chunked and normalized;
- **Connectivity:** Implicit;
- **Completeness:** Mostly fragmented, learners usually capture quick incomplete notes;
- **Context:** Mostly not provided because users focus on writing down the vocabularies and course notes rather than being aware about the context;
- **Heterogeneity:** High heterogeneity because the information is being learned from a variety of sources, so consistency might be low.

3 The Problem

Despite investing effort into collecting vocabulary and study materials, language learners often struggle to turn these resources into effective, contextualized practice.

First, vocabulary frequently remains isolated from real usage. Personal lists tend to store translations and basic information such as part of speech, but rarely show how meanings shift

across contexts. For example, the German verb *"vorstellen"* can mean "to introduce someone", "to imagine something", or "to present an idea", depending on the situation. Without seeing such variations in authentic examples, learners may understand a dictionary meaning but still hesitate to use the word correctly.

Second, learners' language input is scattered across multiple sources: university lectures, grammar PDFs, online lessons, movies, and spontaneous conversations. Notes, vocabulary items, and example sentences from these sources are usually stored separately, making it difficult to reconnect a word to where it was encountered or to quickly access context and explanations that belong together.

Third, review practices are often passive and repetitive. Learners cycle through the same flashcards or word lists, focusing on translation recall rather than using the language in varied, meaningful sentences. Over time, their collections of words, notes, and documents risk becoming static archives instead of dynamic tools that support active communication.

These issues motivate the need for a system that can automatically link a learner's own materials (vocabulary, notes, and grammar texts) and transform them into diverse, context-rich exercises that promote real usage rather than mere memorization.

4 The Solution

4.1 Concept

The system will perform at least the following tasks:

- **Understand:** The system would need to extract and parse the data from different sources, where the system would understand vocabularies, parts of speech, and course notes;
- **Retrieve:** The system needs to retrieve the most relevant examples and grammars' explanations when needed;
- **Generate:** The system would be able to generate new examples for vocabularies, provide grammars' explanations, and create different types of exercises;
- **Track:** The system would need to track the user learning progress and growth.

4.2 User Interface

A visual representation of the data and relations between data makes the most sense for this kind of application. The user should be able to link vocabulary to specific contextual use cases and be guided by the application in an organized way. Instead of having to check multiple sources of data, ex handwritten notes, flashcards etc, the user should now be able to easily find needed vocabulary promptly. Visual methods seem best suited to display lists of vocabulary, try out generated test exercises, and explain possible questions via chat prompt. A web application is also easily available on multiple devices, so users can use it on demand at home, university, or to quickly look up something when, for example, they encounter a new word.

They are also always able to practice with exercises that are generated for their personal experiences/needs. With the interface, they are also able to easily review exercises, find used vocabularies, and understand mistakes made.

4.3 Technical approach

The most obvious technical challenges, and approaches to consider for now are mentioned below even though more challenges might arise over time.

- **Understanding and Structuring::** We will need to figure out how to parse and extract data from hand written notes inside pdf files, and other sources and structure it in a meaningful way. LlamaParse looks promising and worth trying for this challenge.
- **Retrieval and Search:** We will need to experiment with different types of chunking and embedding. Furthermore, we will need to decide how to store the data inside RAG's vector database and the indexing methods. We will also need to decide on the best retrieval method such as semantic search or BM25 or maybe a hybrid approach.
- **Tracking progress:** We need to find the most optimal way to store users' progress, which is helpful for the system to recognize learning progress or knowledge gap of the user when suggesting new vocabularies or exercises.

4.4 Evaluation

A successful AI language tutor can (1) generate a diverse set of exercises, (2) that are relevant and language level appropriate to the user, and (3) accurately distinguishing wrong solutions from correct ones.

4.4.1 Measuring Diversity

We plan on measuring the diversity of generated exercises by two different criteria. First we will measure the distribution of generated exercises over the users vocabulary. This distribution would ideally be uniform to engage the user with every word on their list evenly. For the second diversity criteria, we will use an embedding model to transform the exercises into vector space and then calculate both, the within group and the between group similarity through the pairwise cosine similarity. A group in this context are all exercises that are meant for a word. Ideally the cosine similarity would be small for both, between and within group, suggesting a high diversity in the generated exercise.

4.4.2 Measuring Relevance

To measure the context relevance, we will again use the embedding pipeline from before, but this time we will measure the similarity to the user provided context (notes, texts, etc.). Ideally an exercise for a given word, has a high cosine similarity with related context, suggesting effective generation of relevant exercises. Additionally, we will measure the distribution of words used in the exercises by language level. For german, the Goethe institute publishes lists of words by language level (e.g. A2 list: [1]). A good system will use many words from the list relevant to the users level, and fewer more or less complicated words.

4.4.3 Measuring Solution Grading Performance

To measure how well, the system recognizes wrong and correct solutions, we will curate a dataset of generated exercises with correct and wrong solutions. A successful system will accurately identify wrong and also correct solutions.

4.4.4 Further Measurements

We will further conduct partial system performance measurements by relevant indicators like recall@k and precision@k for context retrieval, or time needed to generate a new exercise.

References

[1] Vera Beiser-Kolb, Andrea Hammann, Monika Reimann, Ulrike Schubert, Frauke van der Werff, and Ulrich Remanofsky. Goethe-institut dtz wortliste. `https://www.goethe.de/resources/files/pdf209/dtz_wortliste.pdf`, 2009. Accessed: 2025-11-17.

Abbreviations

AI Artificial Intelligence