5009 most common words in German according to the Routledge Frequency Dictionary, in both json and csv format

fercho111 Update LICENSE 2308bf4 · 9 months ago

LICENSE	Update LICENSE	9 months ago
README.md	added LICENSE and README with e…	9 months ago
german_vocab_5009.csv	added csv	9 months ago
german_vocab_5009.json	files for usage	9 months ago

README License

Routledge German Frequency Vocabulary Parser

This repository contains a clean, structured dataset extracted from the *Routledge Frequency Dictionary of German*, in both `json` and `csv` format. The goal of the project is to provide a machine-readable version of the 5009 most frequent German words, including part-of-speech tags, example sentences, and English translations, suitable for analysis, Anki deck generation, or other educational and linguistic applications.

Dataset Structure

JSON Format

The `german_vocab_5009.json` file contains a list of 5009 dictionary entries. Each entry has the following structure:

- `id` : Original dictionary entry number.
- `word` : The German word or expression.
- `senses` : A list of senses, where each sense includes:
  - `sense_number` : Number of the sense as it appears in the dictionary.
  - `pos` : Part of speech (e.g., `noun` , `verb` , `adv`).
  - `gender` : Present for nouns (e.g., `der` , `die` , `das`) if available.

    - `translation` : English translation.
    - `example_de` : German example sentence.
    - `example_en` : English translation of the sentence.
- `frequency` : Corpus-based frequency.
- `dispersion` : Measure of how evenly the word appears across text types.

CSV Format

The `german_vocab_5009.csv` file is a flattened version of the JSON. Each **sense** of a word appears as its own row. This format is particularly suitable for data analysis and flashcard tools like Anki.

The columns in the CSV are:

- `id`
- `word`
- `pos`
- `gender` *(if applicable)*
- `translation`
- `example_de`
- `example_en`
- `frequency`
- `dispersion`

For example, a word like *der* with three distinct senses appears as three separate rows in the CSV.

Source Acknowledgement

This dataset is based on content from:

> *A Frequency Dictionary of German: Core Vocabulary for Learners*, by J. Möhring, E. Tschirner, E. Muntschick, and R. Jones (Routledge, 2020)

Disclaimer

This repository provides parsed data from a copyrighted book for **educational, linguistic research, and personal study purposes only**.
No commercial use is intended. If you are the copyright holder and have any concerns or would like this material removed, please contact me directly.

License

See `LICENSE` for details.

Releases

No releases published

Packages

No packages published