Supplementary Material for *Betthupferl:* Data Statement & Annotation Guidelines

VERSION 1.0

Verena Blaschke, Miriam Winkler, Barbara Plank
LMU Munich
{verena.blaschke, b.plank}@lmu.de

May 31, 2025

This is a companion document for the paper

> Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank (2025). "A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation." In *Proc. Interspeech 2025*.

It consists of:

(§1) a data statement,

(§2) a description of the principles according to which we transcribed the audio clips and translated them into Standard German,

(§3) a description of how we annotated differences between the dialectal and Standard German references,

(§4) the guidelines we used for the word-level error analysis, and

(§5) the guidelines we used for the sentence-level automatic speech recognition (ASR) quality judgments.

1 Data statement

This data statement is based on the data statement templates for language datasets by McMillan-Major and Bender (2024) and for speech datasets by Papakyriakopoulos et al. (2023) and (Agnew et al., 2024). It extends the information in sections 3.1 and 3.2 of the paper.

McMillan-Major and Bender, 2024: "A guide for creating and documenting language datasets with data statements" techpolicylab.uw.edu/data-statements
Papakyriakopoulos et al., FAccT 2023: "Augmented datasheets for speech datasets and ethical decision-making"
Agnew et al., ArXiv preprint 2024: "Sound check: Auditing audio datasets"

1.1 Summary and motivation

- Dataset name and version: Betthupferl (version 1.0)
- Dataset curators: Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank
- Dataset citation: Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank (2025). "A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation." In *Proc. Interspeech 2025*
- Data statement version: 1.0
- Sample citation for the data statement: "The data statement (version 1.0) for Blaschke et al. (2025)"

Executive summary Betthupferl is an evaluation dataset for testing how well ASR systems can transcribe audio in three Upper German dialect groups (East Franconian, Bavarian, Swabian). It includes four hours of dialectal speech and half an hour of Standard German speech. Each sentence has a Standard German reference transcription, and (for the dialectal speech) a dialectal reference.

Curation rationale Although there are many non-standard dialects related to Standard German, they are barely represented in ASR datasets. The purpose of *Betthupferl* is to enable studies on how robust ASR systems are towards dialectal variation.

We collaborated with a local broadcasting station (Bayerischer Rundfunk; BR), which has published audio recordings of (mostly dialectal) stories for children. We selected these stories as they contain multiple language varieties (Swabian, East Franconian, different Bavarian dialects, Standard German), have the same genre, and were produced in a high-quality manner.

1.2 Composition

Betthupferl consists of sentence-level audio clips and references.

Linguistic situation and speech characteristics

- Time and place of linguistic activity: The audio clips were recorded in the 2010s and early 2020s.
- Date of data collection: 2024–2025
- Modality: spoken (read speech, edited), written (transcriptions)
- Genre and original intended audience: audio stories for children, broadcast by the Bavarian radio broadcaster Bayerischer Rundfunk (BR)

Dataset statistics Betthupferl includes 273 min (3 788 sentences) of audio data, with sentence-level transcriptions. It contains 32–37 min per administrative region of Bavaria, as well as 32 min in Standard German. Table 2 in Blaschke et al. (2025) provides details such as the distribution across regions, numbers of speakers by region, and average sentence lengths.

Each audio clip corresponds to one sentence, with a mean duration of 4.3 ± 2.9 s. The dataset currently only consists of a test split.

Language users The stories were written by authors with backgrounds in children's literature, screenplays, theatre, and radio. The recordings are by professional speakers who speak the dialect of their respective *Regierungsbezirk* (administrative region) in Bavaria (in the case of the dialect recordings) and Standard German. We include recordings from eight women and fourteen men (perceived gender).

Language varieties We include Standard German spoken in Germany (de-DE) as well as dialects spoken in the administrative regions of Bavaria (see Figure 1 in Blaschke et al. (2025)):

- Lower Franconia: East Franconian (vmf-DE),

- Upper Franconia: East Franconian (vmf-DE), possibly from the transition region with North Bavarian,
- Middle Franconia: East Franconian (vmf-DE), possibly from the transition region with North Bavarian or Swabian,
- Upper Palatinate: North Bavarian (bar-DE), possibly from the transition region with Central Bavarian,
- Lower Bavaria: Central Bavarian (bar-DE), possibly from the transition region with North Bavarian,
- Upper Bavaria: Central Bavarian (bar-DE), possibly from the transition region with South Bavarian or Swabian,
- Swabia: Swabian (swg-DE).

For one story per region, we manually annotate the types of differences between the dialectal and Standard German references (details in §3). The results (aggregated over stories) are in Table 3 of the paper.

Transcriber The transcriber (and dialect-to-standard translator) is a native speaker of German and a Central Bavarian dialect from Upper Bavaria. She is a Master's student in Computational Linguistics in her twenties. She has experience in annotating Bavarian text data for natural language processing (NLP) datasets.

Dataset curators The dataset curators are all native speakers of German. One (=the transcriber) also speaks a Central Bavarian dialect, one an East Franconian dialect, and one a South Bavarian dialect. Three of the curators worked on this dataset as part of their work at an academic NLP research group, and the other two as part of their work at the BR archives.

1.3 Collection and preprocessing

Data source We select stories from BR's *Betthupferl** series. These stories are copyrighted by BR. They are high-quality recordings that were produced and edited by professionals. We selected a subset of stories such that we have a little more than half an hour of audio data for each administrative region and for Standard German (duration measured after preprocessing).

*Bavarian for "bed-time stories"

Preprocessing of the audio data We remove intro/outro segments. The majority of recordings are noise-free. A few audio snippets contain noises like laughter, sighs, other breath sounds or music. If possible, we remove them from the audio clips. If they appear in the middle of sentences, they are left in. However, the noise is not part of the transcription, unless they are part of the story-telling, e.g., like the laughter in *Der hat an Felix obrummt: "Du mi dabatzn… Hahaha!"* (dialect) – *Der hat Felix angebrummt: "Du mich zerquetschen… Hahaha!"* (Standard German; "He growled at Felix: 'You crushing me… Hahaha!'") from the Upper Bavarian episode *Der Zauberer*.

Transcription and translation See §2.

Formatting We convert the audio files from MP3 to WAV. They have a sample rate of 48 kHz (stereo).

The transcriptions are in two folders: `transcriptions_dialect` for the dialectal sentences (and their Standard German translations) and `transcriptions_standard_german` for the Standard German stories and the Standard German sentences that were originally in dialectal stories. We export the transcriptions from the transcription software FOLKER (Schmidt and Schütte, 2010) as FLK files, which we convert to TSV via the software EXMARaLDA (Schmidt, 2004). The following columns of the are important:

1 Speaker code: Male/Female + number + abbreviation of the administrative region (`uba` = Upper Bavaria, `lba` = Lower Bavaria, `upa` = Upper Palatinate, `ufr` = Upper Franconia, `mfr` = Middle Franconia, `lfr` = Lower Franconia, `swg` = Swabia) or `deu` for Standard German. The combination of gender and number is unique for each speaker.

4 Reference transcription

5 Start position in the story's audio clip (in seconds)

6 End position in the story's audio clip (in seconds)

For the dialectal stories, each line with a dialectal transcription is immediately followed by a line with the Standard German translation. For instance, these are the first two lines of `transcriptions_dialect/billy-weltbester-biber-ameise-mundart-oberbayern-1-trans-mw.tsv`:

```
M07-uba   ...   ...   I woit heid eifach amoi mein...          15.5...   21.9...   ...   ...
M07-deu   ...   ...   Ich wollte heute einfach mal meinen...   15.5...   21.9...   ...   ...
```

We also provide code for generating TSV files with an alternative format, which are saved to `data_processed/transcriptions`. Each file in this folder is named according to the pattern `[series title]-[episode title]_[region/language].tsv`. For Standard German sentences, we use the language value `hochdeutschdial` if the sentences were taken from an otherwise dialectal story, and `hochdeutschtest` if they are from Standard German stories. The TSV columns are:

1. Sentence index (zero-indexed)
2. Speaker code (see above)
3. Region/language (as in filename)
4. Dialectal reference (empty for the Standard German audios)
5. Standard German reference
6. Audio duration in seconds
7. File path to audio file

For instance, the first (non-header) line of `data_processed/transcriptions/billy-weltbester-biber-ameise_oberbayern.tsv` is:

```
0   M07   oberbayern   I woit...   Ich wollte...   6.4...   data_processed/audio/billy...
```

Quality control Two other dataset curators (a German speaker, and a speaker of German and East Franconian) checked a random sample of transcripts and corrected a few typos.

1.4 Distribution, use, and maintenance

Distribution Access to the audio data must be granted by Bayerischer Rundfunk on a case-by-case basis due to copyright restrictions (contact Gabriele Wenger-Glemser). We share the transcriptions and translations on Github (`github.com/mainlp/betthupferl`).

Use The dataset has been used once so far: for benchmarking multilingual and German-focused ASR models on Betthupferl (Blaschke et al., 2025). The dataset is designed for ASR. Use of the audio data is granted on a case-by-case basis (see previous paragraph), which also depends on the intended use case.

Maintenance We do *not* plan regular updates to the dataset. If you find errors in the transcriptions, linguistic comparisons, or code, please email Verena Blaschke and/or Barbara Plank.

1.5 Limitations, acknowledgements, and further information

Limitations We only provide one dialectal transcription per sentence. However, as the dialects are not standardized, another transcriber might make other spelling choices. We also only provide one Standard German translation per sentence, although other speakers might translate the sentence slightly differently. In the paper introducing the dataset (Blaschke et al., 2025), we take this into account by conducting qualitative analyses of the ASR hypotheses (section 4.3 of the paper).

The transcriber is a native speaker of only one of the included dialects. However, since the same person transcribed all sentences, the transcription and translation style are uniform across all stories.

Disclosures and acknowledgements There are no conflicts of interest. This research is supported by European Research Council (ERC) Consolidator Grant DIALECT 101043235.

Glossary BR = Bayerischer Rundfunk (public-service radio/TV broadcaster in Bavaria, Germany)

About this data statement A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 3 Schema. The template was prepared by Angelina McMillan-Major and Emily M. Bender and can be found at `http://techpolicylab.uw.edu/data-statements`.

We adapted the structure and questions of the template to better match our dataset, and to integrate parts of the speech-specific data statements by Papakyriakopoulos et al. (2023) and Agnew et al. (2024).

2 Transcription & translation

This section expands on section 3.2 in the paper and the data statement above.

General notes We transcribe each story on a sentence level. For the dialectal sentences, we provide two transcriptions: a dialectal one that closely follows the utterance, and a Standard German transcription. The Standard German sentences are only transcribed in Standard German. Some characters in the dialectal stories speak Standard German instead of the relevant dialect (often to mark them as outsiders or as otherwise socially distinct from the main characters). We remove sentences that are entirely in Standard German from the dialect splits and instead add them to the Standard German split.

For about 30 % of the dialect episodes, BR provided official scripts in either the dialect or Standard German. If available, we use these scripts as the starting point for the transcription. We additionally check for all episodes whether we can access episode descriptions – if so, we use them to inform how we spell character names.

The transcription distinguishes narration from direct speech: direct speech starts and ends with quotations marks. If multiple consecutive sentences are direct speech by one character, each sentence is marked with quotation marks. We use ASCII quotation marks ("…") rather than German quotation marks („…“).

Dialect transcription None of the dialects included in Betthupferl have widely adopted orthographies. We therefore use an ad-hoc spelling style that is based on pronunciation and informed by the grapheme-to-phoneme correspondences in the Standard German orthography. If a dialect-specific word is unclear or unknown to the transcriber, we consult openly available online dialect dictionaries (e.g., the Franconian Dictionary* or the Bavarian Dictionary,** both released and maintained by Bayerische Akademie der Wissenschaften) or local newspaper articles as guidance for finding typical regional spellings.

*Fränkisches Wörterbuch lexhelfer.wbf.badw.de

**Bayerisches Wörterbuch lexhelfer.bwb.badw.de

Translation into Standard German For the translation into Standard German, we try to preserve the syntactic structure of the dialect sentence as much as possible, while keeping the translations as natural-sounding as possible. We translate dialect-specific words into Standard German equivalents, e.g., *Moggele* (Upper Franconia) → *Kälbchen* “calf”. If a word has a cognate in Standard German that is regionally marked and used much less frequently than a non-cognate synonym, we choose the non-cognate: *Buama* (Upper Bavaria) → *Jungen* “boys” (instead of *Buben*). If the transcriber does not know the meaning of a dialect word, we consult dialect dictionaries.

We recommend the following principles for dialect transcriptions:

- Remove non-dialect sentences and noise if possible.
- Verify official spellings of character names for data consistency if available.
- Verify the meaning of dialect-specific words through dialect dictionaries or local sources.

3 Annotation of linguistic differences between dialectal and standard German references

This section corresponds to *3.3. Differences between the transcriptions*. We select one story per administrative region and annotate word-level differences between the dialectal and Standard German references. The annotator is a native speaker of German, with a Master's degree in Computational Linguistics and experience in dialectology.

We align the dialectal and Standard German transcriptions on a word level (allowing gaps). We choose the categories below (Roman numerals) based on the following principles:

- they should be grounded in the linguistics literature on differences between Standard German and Upper German dialects,*
- they should provide a starting point for the word-level ASR error analysis (§4),
- categories should not only contain very few items (we originally started the annotation with a more fine-grained list of categories, which we later merged into larger categories).

*For a brief overview on syntactic differences between Standard German and Bavarian, see Blaschke et al. (2024a,b), for a general introduction to German dialect syntax, see Fleischer (2019).

We annotate the following differences (more than one can apply at a time):

i Phonetic and/or morphological differences reflected in the spelling, e.g., *dahoam* (dialect) – *daheim* (standard) "at home", *aufwach* – *aufwache* "(I) wake up", *bissl* – *bisschen* "(a) bit" (with different diminutive suffixes).

ii Different word choice, e.g., *Moggele* – *Kälbchen* "calf". We include direction adverbs here (*nei* – *rein* "into"), differences in choosing demonstrative pronouns vs. personal pronouns (*des* "this" – *es* "it"), and also use this label for composite words that are partial cognates, but differ in at least one of their content word components. The differences can also concern function word differences (if they are only on the level of single word differences and don't otherwise result in different grammatical structures), e.g., *wia i wieder aufwach* – *als ich wieder aufwache* Word choice differences can be on a word level or a phrase level, but they are only counted once per different construction either way.

iii Different word order, e.g., *… hods da Edi gefrogt* – *… hat Edi sie gefragt* "… Edi asked her" (lit. "had-her the Edi asked; had Edi her asked"). We only count this as one difference per phenomenon.

iv Different word splitting, e.g., *d'Stross* – *die Straße* "the street".

v Determiner before personal names: personal names are generally preceded by the definitive determiner in Upper German dialects (but not in (formal) Standard German), e.g., *da Edi* – *Edi* "(the) Edi".

vi Dropped/fused pronouns in the dialect, e.g., *Kummst mid?* – *Kommst du mit?* "Are you going to join me?" (lit. "Come.2SG (you) with?")

vii Possessive constructions: *ausm Ferdinand seiner Klass* – *aus Ferdinands Klasse* "from Ferdinand's class" (lit. "from the.DAT Ferdinand his class" – "from Ferdinand's class")

viii Different verb-related constructions (counted once per verb):

- Tense differences: *redn kena hod – reden konnte* “could talk” (lit. “talk.INF can.INF has” – “talk.INF could”)
- Auxiliary *tun* “do”: *stehn dud – steht* “is standing” (lit. “stand.INF does” – “stands”)
- Infinitive constructions with/without nominalized infinitives: *Limo zum Hole – Limo zu holen* “to fetch lemonade”

ix Other differences. Examples:

- Words are inserted/omitted for fluency or grammaticality, e.g., particles like *denn* in dialectal questions or additional relativizers (*wo*) in the dialect.
- (Cognate) nouns have different grammatical genders, which is reflected in the inflections of other words (counted once per noun).

4 Word-level error annotation

We compare the model-generated hypotheses to the corresponding Standard German references (section *4.3.2. Error analysis* in the paper). We select the same stories as for the annotation of linguistic differences (§3), and the annotator is the same person. For each word in the Standard German reference, we check if the corresponding word in the hypothesis is the same or different. We discern between the following cases and subcases:

✔ The hypothesis is identical to the reference.

✔ The hypothesis is different, but acceptable:

- The hypothesis is closer to the dialect reference than the Standard German in a way this is acceptable as informal/regional German.
- The hypothesis is normalized to a larger degree than the Standard German reference (e.g., a perfect is turned into a preterite).
- Different (valid) translation of a dialectal term.
- Acceptable spelling variation – e.g., numerals (*15* vs. *fünfzehn*), or plausible alternative spellings of proper nouns (*Lisi* vs. *Liesi*).
- Different, but acceptable, word splitting.
- Punctuation differences that don’t change the meaning of the sentence.

✘ The hypothesis is different and wrong:

- The error is related to wrong word segmentation:
  * One nonsense word instead of several normal ones.
  * Multiple nonsense words instead of one normal one.
  * Many-to-many splitting issue.
- Wrong word (one-to-one correspondence between words in the reference and hypothesis).
- Mangled proper name.
- Correct lemma, but wrong inflection.

- Deletion of a word in the reference.
- Insertion of an additional word into the hypothesis.
- Punctuation or capitalization difference that is not acceptable (results in a different meaning).

We do the above pass first. Then we look at the cases where the dialectal and German references differ (§3), other than phonetically or morphologically (i), and compare the ASR hypothesis to the dialectal and Standard German reference. If the difference is on a (multi-word) phrase level, we compare the entire corresponding phrase. We annotate whether the hypothesis…

✅ … follows the Standard German reference.

☑ … follows the dialectal reference, and the result is acceptable in (informal/regional) German.

☑ … uses an alternative, valid translation of dialect word.

❌ … is nonsensical or ungrammatical.

5 Sentence-level ASR quality annotation

Our ASR quality judgments are inspired by the evaluations of Whisper on Swiss German by Dolev et al. (2024). The annotation instructions we gave to our annotators were in German (§5.1). Here, we additionally provide an English translation (§5.2). The results of these annotations are summarized in section *4.3.1. Human judgments of ASR quality* in the paper.

Dolev et al., VarDial 2024: "Does Whisper understand Swiss German? An automatic, qualitative, and human evaluation."

5.1 Original text (in German)

Wir wollen untersuchen, inwieweit automatische Metriken zur ASR-Bewertung mit unseren Bewertungen korrelieren. Deswegen haben wir für ein paar der Betthupferl-Geschichten hier die Outputs eines der ASR-Systeme gesammelt.

Bitte bewertet die automatisch transkribierten Sätze in der dritten Spalte, indem ihr sie mit den hochdeutschen Referenzen (und wenn nötig mit den Dialektreferenzen) vergleicht. Wir bewerten sowohl den Sinn als auch, wie flüssig der Text klingt, jeweils auf einer Skala von 1 (sehr schlecht) bis 5 (perfekt). Pro transkribierter Geschichte sollte das etwa 10–15 Minuten in Anspruch nehmen.

Bitte nur in das eigene Spreadsheet schauen und sich nicht von den Bewertungen anderer Leute beeinflussen lassen!

Wichtig:

- Jede:r sollte ein **eigenes Spreadsheet** verwenden.
  - Einfach eins der "[name initials here]"-Spreadsheets umbenennen und verwenden – falls es keine freien mehr gibt, bitte das Template-Spreadsheet kopieren und umbenennen. *[link to template]*
- Jede:r sollte die erste Geschichte (Billy/Frosch) annotieren. Dann können wir anschließend diese Annotationen verwenden, um zu vergleichen, wie (un)einig sich die Annotator:innen sind.

- Für die anderen Geschichten bitte diese Tabelle zur Koordination nehmen und markieren, welche Geschichten man gerade annotiert bzw. schon annotiert hat: *[link to overview spreadsheet]*
- Mehrere Annotator:innen pro Geschichte sind okay (bzw. sogar erwünscht, falls es dafür genug Zeit gibt), aber davor sollte zumindest jede Geschichte mindestens einmal annotiert worden sein.

Bedeutung: Hat die automatische Transkription dieselbe Bedeutung wie das Original?

1 – gar nicht: Ganz andere Bedeutung oder unverständlich

2 – kaum: Hat wenige Bedeutungsgemeinsamkeiten mit dem Original

3 – teilweise: Teile des Satzes sind anders oder fehlen

4 – fast: Fast dieselbe Bedeutung; man versteht den Satz auch trotz der Bedeutungsunterschiede im Großen und Ganzen richtig

5 – voll und ganz: Dieselbe Bedeutung, auch was Nuancen anbelangt. Typischerweise identisch zur Referenz bis auf Unterschiede in der Zeichensetzung, Standardisierung ("die Lisa" vs. "Lisa", "ich hab" vs. "ich habe"), Übersetzung von Dialektbegriffen ("Bazi" → "Gauner" vs. "Schlitzohr").

Zusätzlich – falls der Satz zwar inhaltlich der Referenz voll und ganz entspricht, aber (bis auf eventuelle Übersetzungen von Dialektbegriffen) keine wörtliche Transkription ist, bitte in der Spalte "**Anmerkungen**" vermerken. Ansonsten kann diese Spalte leer bleiben. Falls es andere Anmerkungen gibt, können diese in der **Kommentar**spalte vermerkt werden.

Flüssigkeit: Klingt der Text flüssig, natürlich und grammatikalisch richtig? Ist er "gutes Deutsch"?

1 – gar nicht: Gar nicht flüssig, natürlich und/oder grammatikalisch. Inakzeptabel.

2 – ziemlich unnatürlich

3 – merkwürdig: Nicht ganz flüssig/natürlich, klingt merkwürdig

4 – fast: Nur geringe Probleme

5 – voll und ganz: Genau so flüssig, natürlich und grammatikalisch richtig formuliert wie die Referenz

Falls ihr Kommentare zu den Referenzen selbst habt, bitte entsprechend in **Anmerkungen** markieren und in der Kommentarspalte beschreiben.

Vielen Dank für eure Mühe! :)

5.2 English translation

We want to research to what extent automatic ASR quality metrics correlate with our judgments. To this end, we collected the outputs of one of the ASR systems for some of the Betthupferl stories here.

Please rate the automatically transcribed sentences in the third column by comparing them with the Standard German references (and when necessary with the

dialectal references). We rate whether the transcription makes sense as well as how fluent it sounds, both on a scale from 1 (very bad) to 5 (perfect). This should take 10–15 minutes for each transcribed story.

Please only open your own spreadsheet so you're not influenced by the others' ratings!

Important:

- Everybody should use their **own spreadsheet**.
  - Just rename one of the "[name initials here]" spreadsheets umbenennen – if none are left, please copy and rename the template spreadsheet. *[link to template]*
- Everybody should annotate the first story (Billy/frog) so that we can afterwards use these annotations to compare how much the annotators (dis)agree with each other.
- For the other stories, please use this overview table to coordinate annotations and mark which stories you have annotated or are annotating: *[link to overview spreadsheet]*
- Having multiple annotators per story is fine (and even desired if we have enough time), but before that, each story should have already been annotated by at least one person.

Meaning: Does the automatic transcription have the same meaning as the original one?

1 – not at all: Completely different meaning or the transcription doesn't make sense

2 – barely: There are few similarities in meaning with the original

3 – partially: Parts of the sentence are different or missing

4 – almost: Almost the same meaning; the sentence is generally understood correctly despite the differences in meaning

5 – completely: The same meaning, even when it comes to nuances. Typically identical to the reference except for differences in punctuation, normalization ("die Lisa" vs. "Lisa", "ich habe" vs. "ich hab" [presence or absence of determiners before person names, standard verb endings vs. schwa elision]) or the translation of dialect terms ("Bazi" → "Gauner" vs. "Schlitzohr" [different terms for "rascal"]).

Additionally, if the sentence has the exact same meaning as the reference but is not a literal transcription (except for potential translation differences for dialectal terms), please note this in the column "**Notes**". Otherwise, this column can stay empty. If you have other comments, you can add them to the "**Comments**" column.

Fluency: Does the text sound fluent, natural and grammatically correct? Is it "good German"?

1 – not at all: Not at all fluent, natural and/or grammatical. Not acceptable.

2 – rather unnatural

3 – strange: Not entirely fluent/natural, sounds strange

4 – almost: Only minor issues

5 – completely: The phrasing is just as fluent, natural and grammatically correct as that of the reference

If you have comments on the references, please mark them in the **Notes** column and describe this in the **Comments** column.

Thank you a lot for your effort! :)

Literatur

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H. Jiang, Ezra Awumey, and Sauvik Das (2024). "Sound check: Auditing audio datasets."

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank (2024a). "MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10 921–10 938. ELRA and ICCL.

Verena Blaschke, Barbara Kovačić, Siyao Peng, and Barbara Plank (2024b). "MaiBaam annotation guidelines." *Computing Research Repository*, arXiv:2403.05902.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank (2025). "A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation." In *Proc. Interspeech 2025*.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli (2024). "Does Whisper understand Swiss German? An automatic, qualitative, and human evaluation." In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pp. 28–40. Association for Computational Linguistics.

Jürg Fleischer (2019). "Vergleichende Aspekte der deutschen Regionalsprachen: Syntax." In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Deutsch*, pp. 635–664. De Gruyter Mouton.

Angelina McMillan-Major and Emily M. Bender (2024). "A guide for creating and documenting language datasets with data statements." Technical report, University of Washington.

Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke (2023). "Augmented datasheets for speech datasets and ethical decision-making." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, p. 881–904. Association for Computing Machinery.

Thomas Schmidt (2004). "EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache." In *Proceedings of Konvens 2004*. Austrian Society for Artificial Intelligence.

Thomas Schmidt and Wilfried Schütte (2010). "FOLKER: An annotation tool for efficient transcription of natural, multi-party interaction." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).