German Counseling Grounding-Act Corpus (GRACO)

Milena Belosevic
Bielefeld University
milena.belosevic@uni-bielefeld.de

Abstract

We present a corpus of 196 German counseling conversations (ca. 25k turns) between advice seekers and counselors from nine domains. A subset of 11.5k turns was double-annotated with grounding acts (e.g., acknowledgments, repairs), attempts to advance the conversation, success of advancing, and conversation phases. Baseline classification experiments with logistic regression and GBERT-base illustrate the impact of class imbalance in grounding-act classification. For logistic regression, train-only balancing improves Macro-F1 from 0.417 [0.377–0.434] to 0.444 [0.394–0.478]. For GBERT-base, performance remains competitive (Macro-F1 0.481), with balancing yielding comparable results under the same evaluation protocol. Given the scarcity of German corpora of naturally occurring conversations annotated for grounding phenomena, we provide a novel resource for both conversation analysis and natural language processing, facilitating the design of realistic human-language model interactions in German. Code and data are available at https://osf.io/6k275/overview.

Keywords: counseling conversations, grounding acts, language models, classification task

1. Introduction

Grounding acts (GA), such as acknowledgments, repairs, or follow-up questions, play a crucial role in establishing common ground, i.e., the knowledge interlocutors share and continuously update during interactions (Clark and Brennan, 1991). Given their central role in human communication, recent work has increasingly focused on GA in large-scale corpora, either by using pre-trained models or by training models to identify or generate GA (Shaikh et al., 2024, 2025). However, existing resources are almost exclusively English (Mohapatra et al., 2024a) and typically lack systematic annotation of grounding phenomena (Mohapatra et al., 2024b; Jokinen et al., 2024).

To address this gap, we introduce a dataset annotated for multiple layers of grounding in German. We focus on counseling because this domain often involves an asymmetry between an expert, i.e., a professional counselor's perspective, and that of a layperson (the advice seeker). The differences between the problem presentation by the advice seeker and the problem assessment by the counselor can lead to communicative problems (Nothdurft et al., 1994; Best, 2020). Therefore, a frequent use of clarification, repair, and acknowledgment moves is expected, making this domain an ideal setting for studying grounding in naturally occurring interactions.

The dataset is designed for two primary purposes: In NLP applications, the annotated categories enable benchmarking of models for dialogue understanding, classification, and generation of GA, as well as evaluation of conversational grounding. In linguistic research, the dataset can be used to support conversation analysis by providing a systematically annotated resource of German counseling interactions, enabling the investigation of how grounding acts vary across domains, phases of counseling, and speaker roles. Beyond modeling, the dataset also has practical applications in counseling and training scenarios, for example, by helping professionals recognize the grounding strategies employed by advice seekers or by linking grounding acts to phases of the counseling process.

Unlike existing German datasets that rely on other linguistic phenomena, domains, and source types, such as annotating dialogue acts in social media data (Zarisheva and Scheffler, 2015) or speech acts in Wikipedia discussions (Ferschke et al., 2012), our corpus consists of naturally occurring face-to-face, telephone, or video counseling conversations, typically involving two interlocutors and focusing on common ground phenomena. This makes it a rare resource for studying (un)cooperative communication styles in authentic contexts in German. Our contributions are as follows:

- We release **GRACO**, a novel German counseling corpus annotated for grounding acts.
- We adapt and extend existing grounding act schemas designed for English to German counseling dialogues.
- We describe the annotation process and inter-annotator agreement for four main annotation categories (grounding act, conversation phase, attempt to advance the conversation, and success of advancing).
- We provide baseline experiments and distributional analyses to support linguistic and com-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6251–6261
11–16 May 2026.

putational research.

We first situate our work in the context of related research (Section 2) and describe the process of dataset creation (Section 3) and the annotation scheme (Section 4). Section 5 presents baseline experiments, while Sections 6 and 7 summarize the results, discuss limitations, and outline future directions.

2. Related work

Building on earlier work, such as supervised speech act classification of messages in German discussion forums (Bayat et al., 2016), recent studies based on German data have mainly focused on the annotation of dialogue acts and speech acts in domains such as politics (Reinig et al., 2024) or offensive language (Plakidis and Rehm, 2022; Leitner and Rehm, 2025b), using existing text or social media (Twitter) corpora. Similarly, Leitner and Rehm (2025a) conducted large-scale experiments on LLMs as classifiers using existing datasets (mainly Facebook posts and tweets). In contrast, we investigate an under-explored phenomenon in German: the annotation of grounding phenomena in authentic counseling conversations.

While dialogue acts capture the function of an utterance in the dialogue, such as requesting or providing some information (Austin, 1975; Bunt, 1994) and speech acts describe utterances as general actions in language use (e.g., asserting, promising, questioning) (Austin, 1975; Searle, 1969), grounding acts describe how an utterance contributes to the establishment, maintenance, or repair of mutual understanding between interlocutors (Clark and Brennan, 1991). Therefore, by providing a dataset of German GA, we follow similar studies on English data, where grounding acts were explored regarding the question of how frequently the inability to establish common ground (grounding failures) occurs in dialogue (Shaikh et al., 2024).

Our annotation schema is based on several taxonomies of GA (Clark and Brennan, 1991; Traum, 1995; Traum and Hinkelman, 1992; Clark and Schaefer, 1989; Traum, 1992). However, to our knowledge, it remains unclear whether and how they apply to languages beyond English. While our annotation schema (see Section 4.1) adopts the categories, such as acknowledgments (signals of understanding, e.g., *okay*), clarifications used for specification or resolving ambiguity (e.g., *What did you say?*(Purver, 2004), or repair (e.g., *the ma– the husband* addressing troubles in speaking or understanding (Schegloff et al., 1977), we also assign categories, such as backchannels to non-grounding phenomena unless they signal explicit understanding (e.g., explicit answers to yes/no questions). Furthermore, we distinguish between follow-up questions (seeking additional information on a prior utterance) and next-turn utterances. In this way, we provide a more detailed distribution of advancing grounding acts as the dominant class in our dataset.

Source	Number of conv.	Domain
Herzberger 2013	77	Immigration office
Larcher 2007	20	Bank counseling / financial advice
Schröder 1985	6	medical, student, financial, psychological, private, social counseling
Abiri 2022	2	Immigration office
Weber 2014	15	Sales counseling
Marbler 2011	8	Immigration office
Brunner 1996	2	Private counseling
Schubert 2003	11	Financial counseling
DGD database (Dialogstrukturen)	25	Private, medical, student counseling
DGD database (Grundstrukturen)	18	Sales, legal, private, student, health insurance counseling
DGD database (FOLK)	12	Student, private counseling

Table 1: Overview of primary source corpora.

3. Dataset creation

Raw dataset. Our corpus draws on publicly available transcripts of naturally occurring German counseling conversations. Audio or video files accompanying the transcripts are not included in our dataset. We adopt a broad definition of counseling, namely as asymmetric dialogues that require common socio-cultural knowledge by the advice-giver and seeker (Nothdurft et al., 1994). Transcripts that were not considered counseling conversations were excluded from the dataset, such as consultation hours at the university (Boettcher et al., 2005) (in contrast to student advising). The material spans a wide variety of domains (e.g., immigration office, financial, legal, private counseling) and diverse recording types (face-to-face interactions, telephone calls, broadcast recordings), with most conversations being face-to-face. As shown in Table 1, the majority of transcripts come from the immigration office conversations. Dataset properties are described in detail in the dataset card (see Supplementary material).

The dataset is also heterogeneous in terms of conversation length and participant structure, with some transcripts spanning several hundred turns (e.g., financial and purchase counseling) and involving multiple speakers. Moreover, some domains, particularly immigration office dialogues, were clearly overrepresented (45% in the raw data). To achieve a more balanced and coherent resource, we normalized dialects to standard German and applied several filtering steps. Text cleaning included removing pause, segmentation, overlap, and alignment markers (+ ++ . . . //[] < >), breath/noise tags like (*H*), °. . . °, :), metacomments ((laugh)), prosodic markings (e.g. capitalization like *hAUs*), removing leftover punctuation-only tokens, dropping standalone artifact letters (*s g f H ff*) and stray digits, normalizing common fillers to a canonical form: *äh, ähm, hm, hmm*, collapsing repeated whitespace and stripping. Because transcription conventions varied across conversations, long contributions by the same speaker in the source corpora were split

into multiple lines. To harmonize segmentation, we define a turn as one continuous contribution by a single speaker until another speaker takes the floor. Consecutive lines with the same speaker ID (e.g., RG1 ... RG1 ...) were therefore merged. Lines from different speakers, even if they share a role label (e.g., RG1 vs. RG2), are not merged. In the evaluation step, all metrics are computed per merged turn, i.e., in multi-party settings (e.g., RG1/RG2), turns are evaluated independently, and only the coarse role (RG/RS) is used. In total, the raw corpus contained 31,379 speaker lines. After merging, the corpus contained 24,338 turns across 196 conversations. All subsequent steps were performed on these merged turns. The raw, unmerged dataset versions are available in the Supplementary Materials.

Annotated dataset. For annotation, a balanced subset of 52 conversations (11,591 turns) was selected from the raw dataset. Selection followed a cap-floor strategy: Immigration office interactions were limited to $\leq$40% of annotated turns, while smaller domains (e.g., medical, social, legal counseling) were fully included to ensure coverage (medical: 2 conversations, 406 turns; social: 1 conversation, 173 turns; legal: 5 conversations, 581 turns). All instances of rare grounding acts (Repair, Reformulation, Clarification) were included. For majority categories (Next Turn, Acknowledge), we randomly sampled turns, capping at 40% per domain to ensure diversity. Additionally, we covered heterogeneous interaction settings by incorporating face-to-face, telephone, radio, and video conversations (face-to-face: 41; telephone: 7; radio: 1; video/TV: 3). Within each domain, we selected a mix of long, medium, and short conversations to balance the efficiency of annotation with the diversity of interactional settings. This procedure ensures coverage across speaker roles and domains, while avoiding a majority-class bias. For example, within the domain of private counseling, we selected two long coaching sessions (756–1072 turns) and one shorter radio counseling dialogue (150 turns). This ensured both length diversity and modality variation while keeping the domain at $\approx$2k turns. Medium-length conversations from Schröder (1985) were excluded to avoid over-weighting the domain, but they remain part of the raw pool for future extensions. The resulting subset encompasses nine domains, with immigration, student, and private counseling forming the largest categories (see Table 2). Smaller domains, such as medical, legal, and social counseling, are also represented, with at least one conversation each. This strategy provides both breadth (encompassing all domains) and depth (a sufficient number of turns per category), ensuring that the annotated data can serve as a representative benchmark for grounding act analysis across institutional counseling contexts. This subset forms our final dataset, serving as the basis for developing annotation guidelines and conducting the main annotation (Section 4), as well as creating training, development, and test splits for experiments (Section 5). While DGD transcripts are included in the annotated corpus for coverage and linguistic analysis, licensing constraints prevented their use in all model-training experiments reported in Section 5.

Domain	Conv.	Sources	Turns	%
Immigration office	25	Herzberger (2013): 18; Marbler (2011):5; Abiri (2022):2	3800	32.8
Student advising	2	DGD database (Dialogstrukturen): 2	1995	17.2
Private counseling	3	DGD database (FOLK): 2; DGD database (Dialogstrukturen): 1	1978	17.1
Sales counseling	4	DGD database (Grundstrukturen): 2; Weber (2014): 2	1205	10.4
Financial counseling	8	Schubert (2003): 5; Larcher (2007): 2; Schröder (1985): 1	1201	10.3
Legal counseling	5	DGD database (Grundstrukturen): 5	581	5.0
Medical counseling	2	Schröder (1985): 1; DGD database (Dialogstrukturen): 1	406	3.5
Health insurance counseling	2	DGD database (Grundstrukturen): 2	252	2.2
Social counseling	1	Schröder (1985): 1	173	1.5
Total	52	11,591	100.0	

Table 2: Domain distribution (11.5k subset).

4. Data Annotation

Two untrained linguists (German native speakers) annotated the entire 11.5k dataset using the annotation guidelines[1]. Before this main annotation phase, they performed the pilot annotation.

4.1. Annotation scheme

Our annotation scheme encompasses four main categories: *grounding act*, *conversation phase*, *attempt to advance the conversation*, and *success of advancing the conversation*. *Speaker role* (RS: advice seeker and RG: counselor) and *domain* were added by the author based on the information from transcripts and not double-annotated. Additionally, after the annotation and adjudication, the author assigned each double-annotated GA to a more general GA type (advancing, non-advancing, disambiguating, non-grounding) following the existing classifications (see Section 2). Specifically, the following schema was used: advancing (acknowledgment, next-turn, follow-up), non-advancing (reformulation, repair, restart), non-grounding (floorholding, truncated, backchannel, other), or disambiguating (clarification, overresponse) (see also Table 3). Since the category *grounding act type* has a reduced number of labels (four) compared to 12 fine-grained GA, it was used for baseline experiments reported in Section 5.1. The fine-grained categories remain for conversation analysis purposes, but for NLP tasks, we suggest the higher-level grouping. Annotation is performed at the turn level. We treat backchannels (e.g., 'hm', 'mhm', 'ja') used purely as continuers as non-grounding.

[1] See supplementary materials at https://osf.io/6k275/overview.

Utterances that confirm understanding or acceptance of a prior proposal are acknowledgments. Crucially, a direct answer to a prior information-seeking or polar question constitutes next turn, not a backchannel. Table 3 provides definitions and examples (underlined) of annotation categories from the 11.5k dataset. For space reasons, only English translations of all German examples are given in the paper.

4.2. Pilot annotation

In the pilot phase, 903 turns were sampled from the 11.5k dataset. At least one conversation from each domain was included. For longer conversations, excerpts of 50–150 turns were selected from different phases of the interaction, while shorter dialogues were included in full.

The annotators labeled this data independently to identify borderline cases and refine the annotation schema and the guidelines. After refinement, all pilot turns were rechecked and adjudicated by a separate reviewer to ensure alignment with the final scheme and were then merged into the corpus. As a result, the released 11.5k dataset already includes the pilot material in its finalized form. Both for the pilot and main annotation, we report IAA using percent agreement, Cohen's κ (nominal; quadratic-weighted κ for conversation phases), Krippendorff's α, and class-specific PSA (positive specific agreement = 2a/(2a + b + c); NA items are excluded from this calculation).

IAA in the pilot phase was highest for *conversation phase* (κ=0.57, α=0.78), while *attempt to advance* (κ=0.34), *grounding act* (κ=0.53) and *success* (κ=0.43) showed moderate reliability (see Table 4). These results highlighted common sources of confusion, particularly in distinguishing between subtypes of advancing and disambiguating acts, and informed the refinement of the final annotation scheme.

Specifically, for *grounding acts*, annotator 1 uses a narrower inventory (6 labels) than annotator 2 (14 labels), which yields asymmetric confusions: *next turn*→*acknowledgment* (34 vs. 28 in the reverse),*next turn*→ *follow-up* (33 vs. 24), and one-way mappings from *next turn* to finer categories used by annotator 2: *overresponse* (25 vs. 0), *truncated* (22 vs. 0), *reformulation* (13 vs. 0). For conversation phases, the main confusion was between *closing* and *exploration*, indicating different thresholds for when closing begins. For *success*, annotator 1 is more conservative: *no*→*Yes* is 114 vs. 72 in the opposite direction. For *attempt to advance*, a similar pattern holds: *no*→*yes* is 207 vs. 66. Overall, disagreements arise from uneven label granularity across annotators and borderline cases between semantically adjacent categories (e.g., *next turn* vs. *follow-up*/*acknowledgment*, and *closing* vs. *exploration*). We used these findings to tighten boundaries for the fine-grained GA labels, define the onset of closing cues, and calibrate decision rules for *attempt to advance* and *success*.

4.3. Main annotation

The results of the inter-annotator agreement (IAA) for the 11.5k dataset are presented in Table 5. Disagreements were resolved by a separate adjudicator, who did not participate in the pilot and main annotation, to produce the gold standard.

IAA is high overall, with grounding acts (κ = 0.85) and conversation phases (κ = 0.80) showing especially strong consistency. More challenging were the success labels, where agreement dropped to κ = 0.63, reflecting the difficulty of judging uptake in borderline cases. Within grounding acts, frequent categories such as *backchannel* and *next turn* reached very high PSA values (95–94%). In contrast, infrequent repair-related categories (*repair*, *restart*) showed considerably lower reliability. PSA for rare categories, such as *overresponse* and *restart*, is low, which reflects their very small number of instances rather than systematic disagreement. Analysis of double annotations shows systematic disagreement patterns. For *grounding acts*, the most frequent confusions were between *backchannel* and *acknowledgment* (126 vs. 92 cases), *next turn* and *truncated* (119 vs. 78), and *reformulation* vs. *next turn* (85 vs. 51). For *conversation phase*, annotators most often disagreed on whether a turn belonged to *closing* or continued *exploration* (232 cases), while smaller mismatches occurred between *exploration* and *opening* (66 cases). For *attempt to advance*, asymmetries were common, with 684 cases of annotator 1 = *no* vs. annotator 2 = *yes*, and 537 in the opposite direction. Finally, *success* showed the lowest reliability: in 2,790 of 6,786 turns annotated with yes and no (41.1%), annotators diverged on *yes* vs. *no*, nearly evenly split between the two directions (1398 vs. 1392 cases).

4.4. Annotation statistics

Figure 1 summarizes the distribution of grounding acts in the annotated 11.5k subset. The corpus is strongly skewed toward *next turn* (47.5%) and *backchannel* (23.5%). This imbalance is typical of natural counseling dialogues and motivates the release of both natural and balanced data splits (see Section 5).

Conversation phases are also imbalanced: *exploration* dominates (91.9% of all turns), while *closing* (5.8%) and *opening* (2.3%) are comparatively rare. Regarding the binary labels, 58.5% of turns are annotated as *attempt to advance*, and 41.5% as *no attempt*. For *success*, 31.1% of attempts

Category	Definition/Examples
Advancing GA	GA that move the dialogue forward. Includes: *Next Turns* (new task-relevant content), *Acknowledgments* (explicit signals of understanding): RG: So there are intersections. RS: There are exactly.), and *Follow-up questions* (requests for elaboration): RG: This degree programme was in the US. RS: Yes. RG: And did you study here in Germany?).
Disambiguating GA	GA that reduce uncertainty without fully resolving it. Includes: *Clarification requests*: RS: What did you say? RG: I wrote down the calendar week, and *Overresponses* (providing more than asked): RG: You have a room there? Other: Yes, I have a room in the student residence.).
Non-advancing GA	GA that detected grounding failures. Includes: *Reformulations* (self-rephrasing): RS: About eight by nine metres; RG: Yes. RS: About ninety to ninety-five square metres, *Repairs* (corrections of misunderstandings): RG: So, one project is supervision. RS: No, the projects are subject-related.), and *Restarts* (resetting after a breakdown): RG: New application or what? RS: What?. RG: New application).
Non-grounding	Turns that maintain interaction but do not contribute to grounding. Includes fillers (*äh, hm*), backchannels (no explicit uptake):RS: When you do something collaboratively RG: Yes. RS: Then you don't have to do it alone.), floorholders (RG: That's also the question of who is what. It doesn't work, so please say I can't do it. RS: Yes. RG: So... RS: Or you just have to practise it.), truncated/abandoned turns (RS: more my environment in the RG: hmhm RS: allow them to participate in my project), and *other* (greetings, apologies, thanks, farewells).
Conversation phase	Position of a GA in counseling trajectory: *Opening*, *Exploration*, or *Closing*.
Attempt to advance	Binary label: whether a GA could advance the conversation. Purely phatic moves = No.
Success	For advancing/disambiguating GA: whether the attempt to establish common ground was ratified in the subsequent turn and progressed conversation (Yes), failed (No), or undecidable (NA).

Table 3: Definitions and examples (underlined) of annotation categories.

Annotation layer	IAA (pilot phase)
Attempt to advance	Agreement: 69.8%; κ: 0.34; α: 0.66
Conversation phase	Agreement: 91.3%; κ: 0.57; α: 0.78
Grounding act	Agreement: 65.8%; κ: 0.53; α: 0.77
Success	Agreement: 72.2%; κ: 0.43; α: 0.71

Table 4: Inter-annotator agreement (IAA) in the pilot annotation phase.

Annotation layer	IAA main phase
Grounding act	Percent agreement: 89.46% Cohen's κ (nominal): 0.85 PSA: acknowledgment 78.99%, backchannel 95.1%, clarification 69.25%, floorholding 67.08%, follow-up 79.37%, next turn 94.02%, other 97.19%, overresponse 2.9%, reformulation 72.27%, repair 60.76%, restart 50%, truncated 81.54%
Conversation phases	Percent agreement: 96.93% Cohen's κ (nominal): 0.80, weighted κ (quadratic): 0.81 PSA: closing 80.87%, exploration 98.33%, opening 83.06%
Success	Percent agreement: 75.93% Cohen's κ: 0.63, Krippendorff's α: 0.63 PSA: yes 61.81%, no 55.47%
Attempt to advance	Percent agreement: 89.47% Cohen's κ: 0.78 PSA: yes 91.1%, no 87.1%

Table 5: Main annotation IAA.

were successful, 27.5% unsuccessful, and 41.5% remained unresolved (NA).

Role-based distributions highlight the asymmetry between advice seekers (RS) and counselors (RG). RGs produce more *next turns* (52.6% vs. 42.2% for RS) and are more likely to attempt to advance (68.5% vs. 48.0%). RSs, in contrast, contribute proportionally more *backchannels* (30.0% vs. 17.5%) and more often leave success unresolved (51.9% NA vs. 31.5% NA for RG). These contrasts underline the institutional roles in counseling dialogues, where the counselor drives the interaction forward and the advice seeker provides feedback and uptake. Detailed cross-tabulations by domain and conversation phase are provided in the supplementary material.

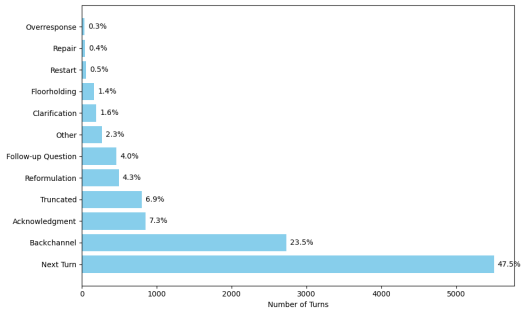

Figure 1: Distribution of grounding acts in the 11.5k annotated subset

5. Case study

Experimental setup. In this section, we evaluate three baselines: majority, TF-IDF + logistic regression, and GBERT-base (Chan et al., 2020), for four-class grounding-act classification (GA-4: advancing, non-advancing, disambiguating, non-grounding (annotation category grounding act type). We also run a context ablation (varying the amount of conversational context). We use conversation-level train/dev/test splits (80/10/10) drawn from the 11.5k annotated corpus.

Balancing. In addition to the natural distribution, we provide a balanced training split to support machine learning experiments on minority classes. We balance training data only via random oversampling with replacement at the turn level. For each class (advancing, non-advancing, disambiguating, non-grounding), we duplicate minority-class turns until each matches the original majority-class count (5,362 turns), yielding a uniform 25% per class in the balanced training split (21,448 turns). Development and test remain unbalanced to reflect real-world skew. Splits are at the conversation level (no conversation appears in more than one split).

Data sizes. Dataset sizes are as follows (in turns): train N= 4,855 (natural) vs 11736 (balanced); dev N= 539; test N= 590. For the experiments reported in this case study, all DGD-derived transcripts were excluded from the training data due to licensing constraints. The splits are therefore based solely on the non-DGD portion of the annotated corpus. DGD-derived conversations remain part of the annotated corpus but are not used for model training.

Metrics. We report Macro-F1 as the primary metric and give 95% confidence intervals via conversation-level bootstrap on the test set (B = 10,000 resamples).

5.1. Classification Experiment

Majority baseline. While the trained models (logistic regression and GBERT) are evaluated on both natural and balanced-train setups, we report the majority baselines on the natural splits only. Baselines use training priors. As trivial baselines, we evaluated a majority-class and a stratified-random predictor on the four-class grounding act task (GA-4). Since the unbalanced test set is skewed towards the advancing class, the majority baseline, which always predicts *advancing*, achieved an accuracy of 0.571 [0.519, 0.620] (conversation-level bootstrap, B=10,000) but only 0.181 Macro-F1 [0.171, 0.191], reflecting its inability to capture minority classes. In contrast, the stratified-random baseline yielded a lower accuracy of 0.475 [0.445, 0.514] but a higher Macro-F1 of 0.280 [0.242, 0.265], indicating more balanced performance across classes despite overall weaker accuracy.

Logistic regression. For the linear baseline, we use TF–IDF Vectorizer over uni/bi-grams (min_df=2, max_df=0.9) with multinomial Logistic Regression (lbfgs, L2, C=1.0, max_iter=1000, random_state=42). Evaluation uses conversation-level natural dev/test, with an additional train-only balanced variant. On the natural main-split test set, using text_ctx0 (current turn only), balanced-train reaches 0.444 Macro-F1 (95% CI: 0.394–0.478), outperforming natural-train at 0.417 (95% CI 0.377–0.434, Δ (difference between new vs. baseline in absolute points) =+0.028). Dev shows a similar gain (from 0.400 to 0.431, 95% CI: 0.367–0.460). Accuracy decreases slightly (from 0.798 to 0.715), consistent with higher recall on minority classes.

Balancing the train-only dataset reduces the tendency of the model to often predict the dominant advancing class when uncertain. Per-class F1 increases for the minority labels (disambiguating +0.095 and non-advancing +0.095) while advancing drops slightly (-0.068) and non-grounding is nearly stable (-0.012). Row-normalized confusions confirm this: The amount of disambiguating turns misclassified as advancing decreases by -45.5 pp, and the share of non-advancing turns misclassified as advancing fell by 29 pp, i.e., the number of minority turns misclassified as the majority class decreases.

Train	N_tr	N_val	N_te	F1_dev	F1_te	Acc_te
Natural (train)	4855	539	590	0.400	0.417	0.798
Balanced (train-only)	11736	539	590	**0.431**	**0.444**	0.715

Table 6: GA-4, TF–IDF+LogReg, current turn only (`text_ctx0`). Train: natural vs balanced; Dev/Test: natural. Metrics: Macro-F1 and Accuracy.

Examples (TF–IDF+LogReg, ctx0).

E1 (Immigration office) *Gold: disambiguating; nat: advancing; bal: disambiguating* RG: So that means you're doing a German language course here in Südstadt now?
E2 (Immigration office) *Gold: non-advancing; nat: advancing; bal: non-advancing* RG: Fourteen days in advance. RS: [...] RG: We always need fourteen days [restart].
E3 (Immigration office) *Gold: advancing; nat: advancing; bal: disambiguating* RG: Your name, please.

The examples illustrate the effect of train-only label balancing. In E1–E2, utterances with clarification cues ("that means") that the natural-train model tends to misclassify as advancing are correctly assigned to the minority labels (disambiguating, non-advancing) by the balanced-train model. By contrast, E3 shows the trade-off: a clearly advancing turn is over-corrected to a minority class. These patterns align with the quantitative results: balancing primarily increases recall for minority labels with a modest reduction in precision for the majority class, yielding a moderate improvement in Macro-F1 on the natural test set.

GBERT. We fine-tune GBERT-base (max length 256; lr 2×10^{-5}, three epochs) on the four-class grounding-act task using text_ctx0 (current turn only) and conversation-level splits. On the natural-train setup, GBERT achieves Macro-F1 = 0.422 on the natural test set (Acc = 0.807), with Dev Macro-F1 = 0.428. Under balanced-train, test Macro-F1 rises to 0.458 (Acc = 0.727) and Dev Macro-F1 improves to 0.431. Thus, balancing raises test Macro-F1 by +0.037 abs while lowering accuracy, consistent with increased recall on minority labels.

Per-class F1 from the test confusions shows that train-only balancing mainly helps the minority classes: both disambiguating (+0.100) and non-advancing (+0.111) improve. Advancing drops

Train	N_{tr}	N_{val}	N_{te}	$F1_{dev}$	$F1_{te}$	Acc_{te}
Natural (train)	4855	539	590	0.428	0.422	0.807
Balanced (train-only)	11736	539	590	0.431	0.458	0.727

Table 7: GA-4, GBERT-base, current turn only (`text_ctx0`). Train: natural vs balanced; Dev/Test: natural. Metrics: Macro-F1 and Accuracy.

slightly (-0.076) and non-grounding remains largely stable (+0.011). Qualitatively, balancing reduces the tendency to map short clarifications or bookkeeping turns to the majority class (advancing), at the cost of a few over-corrections from advancing to minority labels.

Examples (GBERT-base, ctx0).

E1 (Financial counseling) *Gold: Disambiguating; Nat: Advancing; Bal: Disambiguating* RG: Excuse me, what?
E2 (Financial counseling) *Gold: Non-advancing; Nat: Advancing; Bal: Non-advancing* RG: Four months exactly.
E3 (Legal counseling) *Gold: Advancing; Nat: Advancing; Bal: Non-advancing* RG: Do you remember our conversation?

These examples illustrate the dominant error pattern and its mitigation: E1–E2 are typical minority turns that the natural-train model mislabels as advancing, but the balanced-train model corrects. E3 shows overcorrection under balancing.

Per-class F1 confirms the balancing effect for logistic regression: the minority classes (non-advancing and disambiguating), which are not predicted under natural training, receive non-zero F1 scores under balanced training, resulting in an increase in Macro-F1 from 0.417 to 0.444. The advancing class decreases slightly, reflecting the reallocation of predictions away from the majority label. For GBERT-base, performance under natural training already yields strong Macro-F1 (0.481). Under the current data setting, balanced training does not further improve Macro-F1 (0.458), suggesting that the contextualized model is less sensitive to class imbalance than the linear baseline.

In sum, train-only balancing improves Macro-F1 for logistic regression by increasing recall for minority classes and reducing advancing-default predictions. For GBERT-base, performance remains competitive under natural training, indicating that contextualized representations mitigate imbalance effects more effectively than the linear baseline.

5.2. Context-specific ablation

We probe bidirectional context via ctx1 and ctx3. Ctx1 concatenates the previous turn, the current turn, and the next turn. Ctx3 extends this to up to three preceding and up to three following turns. Because following turns are not available at prediction time, this is an oracle setting used to probe signal. At conversation boundaries, windows shorten (no padding). If none exist, the context is empty (i.e., behaves like ctx0).

Logistic regression. We first implemented a multinomial logistic regression classifier (scikit-learn) with TF–IDF features (unigrams and bigrams, min_df=2, max_df=0.9) and L2 regularization (solver=lbfgs, max_iter=1000, random_state=42), varying the amount of surrounding conversational context. The results (Table 10) reveal that performance varies considerably depending on the amount of conversational context available. When using (text_ctx1), the classifier achieves a macro-F1 of 0.29 on the natural splits and 0.34 on the balanced splits, with accuracy scores of 0.62 and 0.54, respectively.

In contrast, extending the context to three turns (text_ctx3) results in a noticeable drop in performance, with Macro-F1 decreasing to 0.22 (natural) and 0.29 (balanced) with accuracy scores of 0.56 and 0.55, indicating that the additional context introduces noise rather than improving discrimination between GA categories. The reason may lie in the fact that bag-of-words features become sparser and noisier, diluting the signal. The difference between balanced and natural splits is particularly evident: while balanced training improves Macro-F1, it also lowers overall accuracy due to stronger penalties on misclassified minority classes.

Confusion matrices show systematic confusions between advancing and non-grounding, as well as between advancing and disambiguating acts. For instance, in the balanced split (ctx1), 63 non-grounding utterances were misclassified as advancing, while 28 advancing turns were misclassified as non-advancing. This reflects the difficulty of distinguishing minimal responses (e.g., ”Yes”) from forward-moving turns.

Examples (LogReg context ablation).

E1 (Immigration office) *Gold: Non-grounding; Nat: Advancing; Bal: Advancing* [prev-1] RG: And the other things [cur] RS: Yes [next+1] RG: I need your passport.
E2 (Financial counseling) *Gold: Advancing; Nat: Non-grounding; Bal: Non-grounding* [prev-1] RG: Did you participate? [cur] RS: Yes. [next+1] rg: And you failed the exam?
E3 (Sales counseling) *Gold: Advancing; Nat: Non-grounding; Bal: Non-grounding* [cur] RS: And my rent was increased by 20 marks. [next+1] RG: Yes.

As shown in the examples, across both natural and balanced splits, misclassifications follow

Label	natural	balanced
advancing	0.843	0.775
non-advancing	0.000	0.095
disambiguating	0.000	0.095
non-grounding	0.824	0.812
Macro-F1	**0.417**	**0.444**

Table 8: Per-class F1 for TF–IDF + LogReg under `text_ctx0`; training: natural vs. balanced (test: natural).

Model	natural	balanced
GBERT-base (Macro-F1)	0.481	0.458

Table 9: GBERT-base Macro-F1 under `text_ctx0`; training: natural vs. balanced (test: natural).

Context	N_{te}	F1 nat	F1 bal	Δ
ctx1	590	0.285	**0.339**	+0.054
ctx3	590	0.223	**0.291**	+0.068
Mean	590	0.254	**0.315**	+0.061

Table 10: Context ablation for TF–IDF+LogReg (`text_ctx1`, `text_ctx3`); Train: natural vs balanced; Test: natural. Metrics: Macro-F1. All rows use the same test set. Δ = balanced - natural (absolute points; positive = better)

Context	N_{te}	F1 nat	F1 bal	Δ
ctx1	590	0.332	0.298	-0.034
ctx3	590	0.344	0.322	-0.022
Mean	590	0.338	0.310	-0.028

Table 11: Context ablation for GBERT-base (`text_ctx1`, `text_ctx3`); Train: natural vs balanced; Test: natural. Metrics: Macro-F1. All rows use the same test set Δ = balanced - natural (absolute points; positive = better).

similar patterns, with consistent difficulties in distinguishing advancing turns from short acknowledgments and non-grounding contributions (E2 and E3). In the natural split, errors cluster around backchannels (E1) and acknowledgments (E2), but also next turns (E3) that were wrongly downplayed to non-grounding GA. In the balanced split, the same errors reappear, indicating that balancing the class distribution improves macro-F1 overall but does not eliminate systematic confusions.

GBERT. The context ablation experiments with GBERT show that performance changes only marginally with more conversational context and balancing does not improve results in this setup. On the natural split, macro-F1 rises slightly from 0.33 (ctx1) to 0.34 (ctx3)(Table 11). A similar trend is visible in the balanced split, where macro-F1 improves from 0.30 (ctx1) to 0.32 (ctx3). This contrasts with the logistic regression results, where additional context decreased performance. Balancing the training set does not increase macro-F1 and substantially lowers accuracy, reflecting that reweighting the class distribution introduces a stronger trade-off without yielding an overall performance gain.

Across both natural and balanced splits, GBERT's misclassifications follow recurring patterns: Minimal confirmations such as short "yes" responses (E1), which in context serve as advancing contributions, are frequently reduced to non-grounding. Clarification requests, such as "where exactly" in E2, are often misinterpreted as advancing, reflecting GBERT's tendency to overestimate progress whenever lexical content is present. Conversely, explicit commitments that clearly advance the task (e.g., "I will bring it tomorrow." in E3) are labeled as non-advancing.

Examples (GBERT context ablation).

E1 (Nat/ctx1, immigration office) *Gold: Advancing; Pred: Non-grounding* [prev-1] RS: Ms Ebert I want [cur] RG: Yes [next+1] RS: to ask you something.
E2 (Nat/ctx3, immigration office) *Gold: Disambiguating; Pred: Advancing* [prev-3] RS: oh [prev-2] RG: I've already added this [prev-1] RS: Where exactly [cur] RG: In the form field 3
E3 (Bal/ctx3, immigration office) *Gold: Advancing; Pred: Non-advancing* [prev-3] RS: zet ha e ein ge like the last letter [prev-2] RG: Yes [prev-1] RS: Okay [cur] RG: I will bring it tomorrow.

These examples show that, even with additional context, GBERT still struggles to capture subtle pragmatic distinctions between acknowledgments, clarifications, and genuine progress markers.

6. Conclusions

We presented GRACO, a novel German counseling corpus comprising 25k turns across nine domains, with an 11.5k subset annotated for grounding act type, conversation phase, attempt to advance, and success. Inter-annotator agreement confirms the reliability of the annotation scheme, with high consistency for conversation phases and grounding acts, and moderate but informative agreement for attempt and success. Baseline experiments with LogReg and GBERT demonstrate that balancing

training data improves Macro-F1 for logistic regression by raising performance on minority classes. For GBERT, performance under balanced training remains comparable but does not exceed natural training.

Together, these findings establish GRACO as a resource for both linguistic research on German grounding phenomena and computational modeling, supporting future work on classification, domain transfer, and generative dialogue tasks, such as next-turn prediction or response generation in advisory dialogues. Initial exploratory trials with generative LLMs (not reported here) indicated that such tasks require more context and careful fine-tuning to produce stable results. We therefore leave a systematic evaluation of generative approaches for future work.

We also plan to extend the annotation to further portions of the raw pool (25k turns) to broaden domain coverage and enable larger-scale evaluations.

7. Limitations

While GRACO provides a novel, large-scale corpus of German counseling dialogues with GA annotation, several limitations remain. First, the data are drawn from publicly available transcripts, which means that not all counseling domains and interaction types are covered. Furthermore, we acknowledge that the raw pool is skewed toward the immigration office and student advising. To mitigate this, we used a cap–floor sampling strategy: no single domain exceeds 40% of the annotated set, and all smaller domains were fully included. While the dataset isn't fully balanced, it ensures diversity across nine domains.

Second, the annotation scheme, while reliable, required interpretive judgments, and moderate agreement on the *attempt to advance* and *success* layers reflects the difficulty of consistently coding borderline cases.

Third, the released dataset does not include audio or video, which may omit prosodic or multimodal cues that are relevant for grounding. However, the choice of transcripts ensures that GRACO is fully anonymized. The released annotation layer and metadata are shareable under an open license, while the underlying source corpora remain subject to their respective licensing conditions. We see this as a foundation that future multimodal work can build on, rather than a complete representation of all grounding phenomena.

Fourth, the annotations were carried out by only two annotators, which allows measurement of pairwise reliability but not broader multi-annotator consensus. As with any manual annotation, potential annotator bias cannot be fully excluded. Borderline cases may be resolved differently depending on individual interpretation, and the use of only two annotators limits the diversity of perspectives. These risks were mitigated through iterative guideline refinement and adjudication (including adjudication guidelines), but they remain a factor to consider in downstream use.

An IAA for *success* (κ=0.63) reflects moderate reliability. We see this annotation category as exploratory. It is included for researchers interested in uptake, but we advise treating it with caution in modeling.

Finally, baseline experiments are limited to text-based classification and do not yet explore generative modeling. Additionally, due to licensing constraints of certain source corpora (DGD), only a subset of the annotated data was used for model training in the reported experiments.

These constraints highlight opportunities for expanding the corpus and refining the annotation in future work.

8. Ethics statement

All data in GRACO are derived from publicly available counseling transcripts that have already undergone anonymization by the original providers. During preprocessing, we removed remaining identifiers such as names, addresses, and artifact letters, and normalized fillers. The dataset, therefore, contains no personal or sensitive information that could be linked to individuals. We acknowledge that counseling dialogues involve sensitive domains, and care was taken to respect participant privacy by excluding audio and video recordings and releasing only de-identified text. The released annotation layer and metadata are distributed under an open license (CC-BY) to encourage reproducible research, with the expectation that they will be used for scientific purposes such as linguistic analysis and computational modeling of grounding, not for applications that might compromise counselor-client confidentiality. The underlying source corpora remain subject to their respective licenses and access conditions.

9. References

Sara Abiri. 2022. *Beratung in interkultureller Kommunikation*. Ph.D. thesis, Universität Hamburg.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Berken Bayat, Christopher Krauss, Agathe Merceron, and Stefan Arbanowski. 2016. Supervised speech act classification of messages in german online discussions. In *FLAIRS*, pages 204–209.

Laura Best. 2020. *Nähe und Distanz in der Beratung*. Soziale Arbeit als Wohlfahrtsproduktion. Springer, Wiesbaden.

Wolfgang Boettcher, Anika Limburg, Dorothee Meer, and Vera Zegers. 2005. *„Ich komm (0) weil ich wohl etwas das thema meiner Hausarbeit etwas verfehlt habe,"–Sprechstundengespräche an der Hochschule. Ein Transkriptband*. Verlag für Gesprächsforschung.

Evald Johannes Brunner. 1996. *Grundfragen der Familientherapie: Systemische Theorie und Methodologie*. Springer, Berlin.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association, Washington.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cogn. Sci.*, 13(2):259–294.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France. Association for Computational Linguistics.

Gesine Herzberger. 2013. *Das sprachliche und kommunikative Verhalten von Behördenmitarbeitern – Agenten-Klienten-Gespräche in einer Ausländerbehörde*. Ph.D. thesis, Univ. Würzburg.

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards harnessing large language models for comprehension of conversational grounding.

Sylvia Bendel Larcher. 2007. *Sprachliche Individualität in der Institution: Telefongespräche in der Bank und ihre individuelle Gestaltung*. Narr.

Elena Leitner and Georg Rehm. 2025a. Exploring the limits of LLMs for german text classification: Prompting and fine-tuning strategies across small and medium-sized datasets. *Journal for Language Technology and Computational Linguistics*, 38(2):1–12.

Elena Leitner and Georg Rehm. 2025b. Exploring the limits of LLMs in German text classification: Prompting and fine-tuning strategies across small and medium-sized datasets. Presented at the LLM-Fails Workshop, IDS Mannheim. Accessed: 2025-04-26.

Marlies Marbler. 2011. *Verständigungsorientiertes Sprachhandeln in interkulturellen behördlichen Kommunikationssituationen/vorgelegt von Marlies Marbler*. Ph.D. thesis, Karl-Franzens-Universität Graz.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. ELRA and ICCL.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024b. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Werner Nothdurft, Ulrich Reitemeier, and Peter Schröder. 1994. *Beratungsgespräche: Analyse asymmetrischer Dialoge*, volume 61. Narr.

Melina Plakidis and Georg Rehm. 2022. A dataset of offensive German language tweets annotated for speech acts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4799–4807, Marseille, France. European Language Resources Association.

Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London King's College.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Peter Schröder. 1985. *Beratungsgespräche. Ein kommentierter Textband*. Narr, Tübingen.

Thomas Schubert. 2003. *Wissenstransfer im telefonischen Beratungsgespräch*. Ph.D. thesis, Martin-Luther-Universität Halle-Wittenberg.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

David R Traum. 1992. A speech acts approach to grounding in conversation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)*, pages 137–140.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Comput. Intell.*, 8(3):575–599.

David Rood Traum. 1995. *A computational theory of grounding in natural language conversation*. University of Rochester.

Peter Weber. 2014. *Verkaufsgespräche im Gartencenter und in der Schule. Ein Transkriptband*. Verlag für Gesprächsforschung.

Elina Zarisheva and Tatjana Scheffler. 2015. Dialog act annotation for Twitter conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 114–123, Prague, Czech Republic. Association for Computational Linguistics.