# Question and Response Dynamics in Public Service Encounters

**Wassiliki Siskou**[1,2], **Ingrid Espinoza**[3], **Laurin Friedrich**[2], **Steffen Eckhard**[2], **Annette Hautli-Janisz**[1]
[1]University of Passau [2]Zeppelin University [3]University of Konstanz
firstname.lastname@{uni-konstanz.de|uni-passau.de|zu.de}

**Abstract**

When deciding on social welfare benefits, street-level bureaucrats wield significant discretionary power over citizens. One of the key instruments of this power lies in the questioning patterns that control the conversational agenda in face-to-face encounters. In turn, the citizens' responses show how they navigate these conversational constraints, for instance by answering directly or through more evasive strategies. To shed light on the power dynamics inherent in these encounters, we provide over 200 verbatim transcripts of authentic conversations in German between street-level bureaucrats and citizens, as well as a fully annotated dataset of all question-response pairs extracted from these conversations. We also present PSE v2.0, which is double the size of the only previously available corpus of spoken interactions in street-level bureaucracy.

**Keywords:** Public Service Encounters, verbatim transcripts, question-response pairs

## 1. Introduction

Institutional face-to-face interactions between representatives of the state and citizens are highly under-represented in conversational corpus linguistics, despite their societal relevance as the one place where citizens and state directly interact. These encounters occur for example in job centers and social welfare offices and are characterized by power asymmetries and information-seeking protocols that differ from everyday-conversations (Thornborrow, 2002). Despite their social significance, publicly available corpora documenting such interactions remain limited. The only publicly available corpus in German is PSE v1.0 (Espinoza et al., 2024), with 106 conversations and about half the size of the dataset presented in this paper. Since its initial release the corpus has served as a resource for research on institutional discourse (Frenzel and Hautli-Janisz, 2025; Friedrich and Eckhard, 2025; Eckhard and Friedrich, 2025) and has been downloaded over 4,000 times. However, the size of PSE v1.0 constrains the scope of quantitative analyses, particularly for lower-frequency phenomena.

Our contribution in this paper is twofold: we present PSE v2.0, a substantially expanded and enriched version of the corpus to tackle this limitation. All data is made publicly available in multiple formats (CSV, TXT, XML) to accommodate different research workflows. Additionally, we provide annotations of all 12,355 question-and-response sequences in the corpus – these sequences constitute almost 20% of all utterances in the data. Questions, which are the primary interactional mechanism through which information is exchanged in institutional dialogue, are classified by syntactic type in PSE v2.0, responses are categorized by their pragmatic function.

## 2. Background

### 2.1. Public Service Encounters

Public Service Encounters (PSEs) are direct, often face-to-face, interactions between citizens and government officials (street-level bureaucrats) where public services are delivered, such as welfare, education, or healthcare. They are considered the "ground floor of government", where policies are applied, and social equity is negotiated Eckhard and Friedrich (2022, p. 3). Street-level bureaucrats therefore hold the responsibility to apply laws and decrees to the personal situations of individuals (Bruhn and Ekström, 2017, p. 211) while upholding principles of fairness and social equity (Raaphorst, 2021; Sumra, 2019; Zacka, 2019, p. 454). In particular, their decisions ultimately define how government support is applied to individual cases for which policies can only provide abstract and vague guidelines (Hupe and Hill, 2007; Lipsky, 2010; Maynard-Moody and Musheno, 2000).
Recent research shows that the language of bureaucrats in these encounters affects citizen satisfaction (Eckhard and Friedrich, 2022; Eckhard et al., 2022). But analyzing the actual communication in these encounters is still in its infant stages, primarily due to data availability. As shown in Espinoza et al. (2024), data protection is paramount in a setting where citizens disclose personal information and bureaucrats are being observed at work, making institutional agreements to record in-person meetings highly time-consuming. PSE v1.0 has paved the way for collecting such data for German, with PSE v2.0 now significantly extending the empirical basis for systematic studies of public service encounters. This allows the computational social science community to conduct large-scale systematic studies on the linguistic properties of



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2848–2855
11–16 May 2026. 

these encounters, as for instance regarding the usage of plain versus complex language (Siskou et al., 2022) and the effect of small talk (Frenzel and Hautli-Janisz, 2025).

### 2.2. Question and response type annotation

**Questions** Corpus work on manual question annotation shows that the phenomenon is challenging to grasp and also depends on the genre. The Switchboard corpus (Jurafsky et al., 1997; Calhoun et al., 2010) encodes information-seeking questions based on their syntactic property (e.g. yes-no-questions, wh-questions) rather than pragmatic function, similar to the MapTask coding scheme (Carletta et al., 1997). Stivers and Enfield (2010) propose a cross-linguistic categorization of questions into seven different social actions, Hautli-Janisz et al. (2022) distinguish between four types of questions for broadcast political debate analysis. Siskou and Espinoza (2024) employ a syntactically-driven annotation scheme attributing questions to 'open' (for wh-questions), 'closed' (for alternative and polar) and 'other' to questions in English parole hearing interviews, using an LLM for annotation. In the present paper we pursue a similar approach adding tag questions, as those are frequent in the dataset at hand, and employing GPT-5 as annotator. As with any large-scale corpus linguistic task, the field is confronted with "a trade-off between usefulness and ease or consistency of coding" (Carletta et al., 1997, p. 15), in particular for a phenomenon as vague and complex as questions (and responses).

**Responses** The variety of question annotations is mirrored in the number of available response taxonomies: Stivers and Enfield (2010) use the four categories of 'Non-response', 'Nonanswer response', 'Answer', and 'Can't determine' (p. 2624) for spontaneous, naturally occurring conversation, showing some overlap with the six-way distinction of Berninger and Garvey (1981) with 'Possible answers', 'Indirect answers', 'Confessions of Ignorance', 'Clarification requests', 'Evasive replies' and 'Miscellaneous' (targeted at nursery school child conversation). Łupkowski and Ginzburg (2017) propose a taxonomy for responses that answer a query with a query, for instance 'Requests for underlying motivation' or 'I ignore your question' – the latter being identical to Berninger and Garvey (1981)'s 'Confession of ignorance' type. For task-based interactions and informal conversations, Ginzburg et al. (2019) propose a two-way distinctions of responses into 'Answers' and 'Non-answers', with fine-grained subcategories (e.g. 'Direct' versus 'Indirect answers', 'Clarification requests', 'Change-the-topic response' and 'Ignore' for non-answers). Thomas et al. (2024) use a hierarchical taxonomy with three higher-level categories ('Clear Reply', 'Ambivalent Reply' and 'Clear Non-reply') and nine subcategories to annotate response clarity in question-and-response pairs from political interviews. In the present paper we use a set of response categories that crosscuts previous work and captures well the discourse dynamics in public service encounters.

## 3. PSE v2.0

With PSE v2.0 we present a substantial expansion of the original Public Service Encounters dataset, doubling the number of transcripts in the final dataset. The dataset is available under the CC-BY 4.0 license at Harvard Dataverse[1]. For data collection, transcription, anonymization and data storage, we follow the exact same protocol described in the PSE v1.0 documentation, thus ensuring consistency and comparability. Table 1 presents a comparative overview of the corpus statistics across both corpora. The total number of transcripts between v1.0 and v2.0 is increased from 106 to 232, representing a growth of 119%. The token count is increased by 307,793 tokens, an additional 71% compared to PSE v1.0. The total number of utterances exhibits an increase from 31,341 to 57,119 (82% growth).

| Metric | v1.0 | v2.0 | Growth |
|---|---|---|---|
| Conversations | 106 | 232 | +119% |
| Tokens | 433,780 | 740,973 | +71% |
| Utterances | 31,451 | 57,119 | +82% |
| Sentences | 59,410 | 115,333 | +94% |
| Questions | 6,033 | 12,355 | +104% |

Table 1: Comparative statistics for corpus versions

While the original corpus primarily contains transcripts from face-to-face interactions from job centers all over Germany, PSE v2.0 adds recordings from tax offices, registration offices as well as more conversations from job centers and social welfare offices. v2.0 therefore captures a wider range of benefit-related consultations and case management interactions.

Regarding the question-response distribution patterns, out of 57,119 utterances, 11,181 (19.57%) contain at least one question. In PSE v2.0, as in spontaneous natural speech in general, the dialogues frequently contain self-repairs, hesitations, pauses and incomplete fragments due to overlapping talk and interruptions from other interlocutors. In total, 37.54% of the utterances contain at least one interruption (indicated by the use of "/" or "//" in

[1] https://doi.org/10.7910/DVN/Y8YU2O

the transcript). Regarding questions, the 115,333 sentences in v2.0 feature 11,833 complete (uninterrupted) questions, and an additional 522 interrupted questions where the speaker is cut off by a different speaker before completing the question. Those interrupted questions are included in our dataset by reconstructing their fragmented units.

## 4. Annotation

### 4.1. Question Types

Questions function as the central motor for information exchange in administrative settings. State officials elicit personal data, clarify citizen circumstances, or verify eligibility criteria (e.g., for welfare benefits), while citizens may seek guidance in a specific bureaucratic procedure, request clarifications, or even challenge the official's decision. While under different circumstances, questions may be considered as a mere conversational formality, in institutional dialogue they reflect and constitute the inherent power dynamics in the interaction.

To systematically analyze the interrogative structures in our dataset in more detail, we adopt a five-category annotation scheme, based on the syntactic properties of questions. Each question (complete or interrupted) is classified into one of the following categories:

**Polar Question** for questions that elicit a yes/no response, typically characterized by a subject-verb inversion in German.

**Tag Question** for declarative statements that are followed by a brief interrogative tag to seek confirmation.

**Alternative Question** for questions that present two or more explicit options to choose from, typically connected by *oder* ('or').

**Wh-Question** for information-seeking questions that are introduced by wh-words like *wer* ('who'), *wo* ('where'), *wann* ('when'), *was* ('what'), *wie* ('how'), and *warum* ('why').

**Other** for questions that do not fit the above categories. These usually include echo questions and ambiguous cases where the interrogative function is unclear.

In order to test the applicability of automatic annotation, one expert linguist manually annotates 747 questions (6%), comprising 219 polar questions (~29.32%), 141 tag-questions (~18.88%), 130 wh-questions (~17.40%), 123 alternative questions (~16.47%) and 134 'other' questions (~17.94%). Most of the 'other' question instances are either assertive sentences that are mistakenly transcribed as questions or ungrammatical sentences that could not be parsed correctly.

We then compare *LiAnS* (Gold et al., 2015), a rule-based classification algorithm that relies on hand-crafted linguistic features tailored to the morphosyntactic features of questions in German, with an LLM, namely GPT-5, for annotation, because recent studies (Mens et al., 2023; Gilardi et al., 2023; Siskou and Espinoza, 2024) indicate good performance of LLMs in annotating datasets[2]. Table 2 presents the performance metrics achieved by LiAnS and GPT-5 for each question type. The performance of both systems varies depending on the question type to be annotated (Macro avg F1 between .47 and .86), but overall, GPT-5 yields a higher performance on the task. We attribute the annotation difficulties to the characteristics of face-to-face spontaneous speech that for instance contains ungrammatical sentences, hesitations, disfluencies, filled pauses and reformulations. In the end, based on the comparative results, we employ GPT-5 for the question type annotation in PSE v2.0.

| | LiAnS | | | GPT-5 | | |
|---|---|---|---|---|---|---|
| **Class** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Alternative | .52 | 1.0 | .69 | .81 | .91 | .85 |
| Polar | .79 | .34 | .47 | .74 | .74 | .74 |
| Tag | .95 | .74 | .84 | .74 | .84 | .79 |
| Other | .44 | .51 | .47 | .58 | .40 | .47 |
| Wh | .78 | .90 | .84 | .84 | .88 | .86 |
| Accuracy | | | .65 | | | .75 |
| Macro avg | .70 | .70 | .66 | .74 | .75 | .74 |
| Weighted avg | .71 | .65 | .64 | .74 | .75 | .74 |

Table 2: Question classification performance showing precision (P), recall (R) and F1-score (F1) for each question type

### 4.2. Response Types

Due to the inherent power asymmetry and role obligations in public service encounters, this data genre is predisposed to contain higher rates of adequate responses compared to genres like political interviews (Thomas et al., 2024). This is due to two factors: On the one hand, street-level bureaucrats are professionally obliged to provide the information requested. On the other hand, citizens answer questions properly due to their dependence on the officials discretionary power. An extensive manual inspection of the transcripts shows that, interestingly, none of the parties always provides clear answers. We therefore capture the pragmatic function and completeness of the responses using the following six-category annotation scheme:

**Proper Answer** for responses that directly or indirectly address the question and provide relevant and complete information.

**Non-Answer** for responses that fail to provide the

[2]See Appendix A.1 for the detailed prompt for the question classification.

requested information. This includes hedged responses, topic shifts, irrelevant responses, as well as evasive responses.
**No Response** for cases where the addressee does not give any verbal response to the question.
**Response with Question** for responses including questions that ask for e.g., clarification.
**Interrupted Response** for responses where the speaker begins to answer but is cut off by another speaker before completing the utterance.
**Other** for responses that do not fit into one of the above categories. These include ambiguous cases or ungrammatical sentences, parallel to patterns in the 'other' class in the question annotation.

For evaluating the classification performance in response types, the linguistic expert manually annotates the 747 responses of the previously annotated questions using this annotation scheme. In total 415 of the questions (~55.5%) are fully answered by providing the requested information, while in 66 (~8.8%) cases the response are either off-topic or evasive. 112 (~15.39%) responses are interrupted before the response utterance could have been completed. In 86 cases, the respondent poses a counter question, either requesting clarification or asking for further information.

Given the high contextual load in response annotation, we use GPT-5 for the annotation[3]. As shown in Table 3, the model achieves an accuracy of .66, with again varied performance among categories. While the model performs well on proper answers (.79), it struggles with Non-Answers (.51) and responses of the 'Other' category (.21). Despite these differences in performance, we use GPT-5 for the annotation of the complete dataset.

| Class | P | R | F1 |
|---|---|---|---|
| Interrupted Response | .53 | .42 | .47 |
| No Response | .73 | 1.00 | .84 |
| Non-Answer | .45 | .59 | .51 |
| Other | .19 | .23 | .21 |
| Proper Answer | .78 | .80 | .79 |
| Response with Question | .88 | .67 | .76 |
| Accuracy | | | .66 |
| Macro avg | .59 | .62 | .60 |
| Weighted avg | .68 | .66 | .67 |

Table 3: GPT-5 response classification performance showing precision (P), recall (R) and F1-score (F1) for each question type.

[3]See Appendix A.2 for the detailed prompt used for response classification.

## 5. Question-Response Dynamics

The conversational dynamics in PSEs are highly influenced by the question types, with closed questions (tag, polar) restricting the response options compared to open wh-questions. Table 4 and 5 show the overall distribution of question and response types in PSE v2.0. Tag and polar questions are the most frequent question categories in the dataset (61%), while open-ended wh-question constitute only 17.2% of the questions in the dataset. Alternative questions seem to occur very rarely in PSEs. Response pattern show that questions are predominantly answered properly (54.2%). Nevertheless, a substantial portion (11.3%) of the questions are being replied to with a non-answer, interrupted (10.7%) or counter-questioned (8.4%).

| Question Type | n | % |
|---|---|---|
| Tag questions | 4,059 | 32.8 |
| Polar questions | 3,594 | 29.1 |
| Other questions | 2,195 | 17.7 |
| Wh questions | 2,127 | 17.2 |
| Alternative questions | 378 | 3.1 |

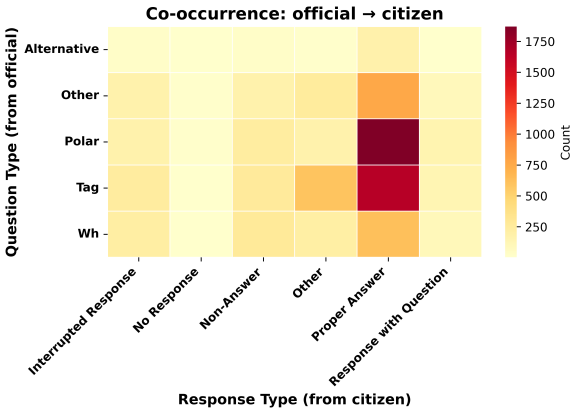
Table 4: Distribution of question types.

| Response Type | n | % |
|---|---|---|
| Proper Answer | 6,695 | 54.2 |
| Other | 1,779 | 14.4 |
| Non-Answer | 1,398 | 11.3 |
| Interrupted Response | 1,326 | 10.7 |
| Response with Question | 1,043 | 8.4 |
| No Response | 74 | 0.6 |

Table 5: Distribution of response types.

The heatmaps in Figure 1 illustrate how the role of each speaker shapes the question-response dynamics in PSEs. The upper heatmap (1a) shows that while the bureaucrats ask significantly more questions than citizens, both parties predominantly receive proper answers, particularly to tag and polar questions. The fact that citizens do ask substantially fewer questions underscores the conversational asymmetry in institutional dialogue and the bureaucrats' authority over the discursive agenda. When citizens do ask questions (lower heatmap (1b)), the response patterns of the bureaucrats yield a more varied picture: The officials produce notable rates of non-answers and responses with counter-questions. These findings indicate that citizens do have restricted response strategies, deciding to provide the requested information consistently.

(a) Official→citizen interactions




(b) Citizen→official interactions

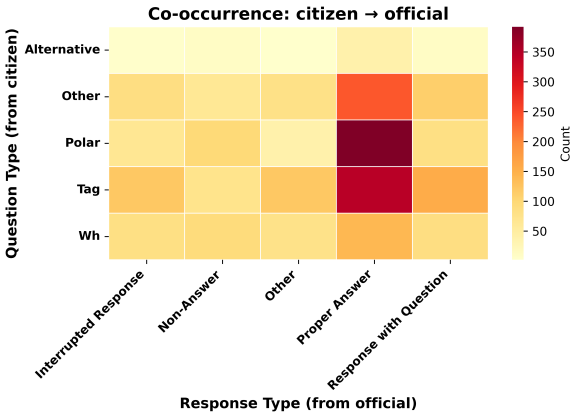



Figure 1: Question-response dynamics

## 6. Summary

PSE v2.0 is a substantial extension of previous empirical work on German public service encounters: We not only provide a corpus double the size of previous available datasets, but we also offer a linguistic annotation of a core pragmatic property in public service encounters, namely questions and responses. We empirically show that officials tend to control the question-response sequences, while citizens take a more responsive role in institutional dialogue.

## Ethics Statement

Given the sensitive nature of this data, we underwent a rigid ethics approval process at the project's institution. We want to emphasize that all participants provided their explicit consent for the recording of their conversations. All transcripts were anonymized manually and rigorously by removing any personal identifier (e.g. names, addresses, date of birth, identification numbers, locations, etc.), replacing them with generic placeholders. While our strict commitment to data protection limited the total number of transcripts collected, we prioritized the participants' privacy over corpus size.

## Acknowledgements

We would like to express our gratitude to the agencies and each official and citizen who participated in the project. We also thank our student assistants who were involved in data collection, management and anonymisation.
The work reported on in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379 as part of the project "Inequality in Street-level Bureaucracy: Linguistic Analysis of Public Service Encounters" at the University of Konstanz.

## 7. Bibliographical References

Ginger Berninger and Catherine Garvey. 1981. Relevant replies to questions: Answers versus evasions. *Journal of Psycholinguistic Research*, 10(4):403–420.

Anders Bruhn and Mats Ekström. 2017. Towards a Multi-Level Approach on Frontline Interactions in the Public Sector: Institutional Transformations and the Dynamics of Real-time Interactions. *Social and Policy Administration*, 51(1).

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Steffen Eckhard and Laurin Friedrich. 2022. Linguistic Features of Public Service Encounters: How Spoken Administrative Language Affects Citizen Satisfaction. *Journal of Public Administration Research and Theory*.

Steffen Eckhard and Laurin Friedrich. 2025. The language of public encounters: Computational measures of complexity and emotionality in spoken bureaucratic communication. *Social Policy & Administration*.

Steffen Eckhard, Laurin Friedrich, Annette Hautli-Janisz, Vanessa Mueden, and Ingrid Espinoza. 2022. A Taxonomy of Administrative Language in Public Service Encounters. *International Public Management Journal*.

Ingrid Espinoza, Steffen Frenzel, Laurin Friedrich, Wassiliki Siskou, Steffen Eckhard, and Annette Hautli-Janisz. 2024. PSE v1.0: The first open access corpus of public service encounters. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13315–13320, Torino, Italia. ELRA and ICCL.

Steffen Frenzel and Annette Hautli-Janisz. 2025. Identifying small talk in natural conversations. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 272–277, Albuquerque, New Mexico. Association for Computational Linguistics.

Laurin Friedrich and Steffen Eckhard. 2025. Inequality in frontline communication: bureaucrats talk differently to men and women. *Journal of Public Administration Research and Theory*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, and Paweł Łupkowski. 2019. Characterizing the response space of questions: a corpus study for English and Polish. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330, Stockholm, Sweden. Association for Computational Linguistics.

Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158. _eprint: https://academic.oup.com/dsh/article-pdf/32/1/141/11046544/fqv033.pdf.

Annette Hautli-Janisz, Katarzyna Budzynska, Conor McKillop, Brian Plüss, Valentin Gold, and Chris Reed. 2022. Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

Peter Hupe and Michael Hill. 2007. Street-Level Bureaucracy and Public Accountability. *Public Administration*, 85(2).

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation – coders manual, draft 13. Technical report, University of Colorado Institute of Cognitive Science.

Michael Lipsky. 2010. Street-Level Bureaucracy: Dilemmas of the Individual in Public Service. *American Political Science Association*, 76(1).

Steven Maynard-Moody and Michael Musheno. 2000. State Agent or Citizen Agent: Two Narratives of Discretion. *Journal of Public Administration Research and Theory: J-PART*, 10(2).

Gaël Le Mens, Balázs Kovács, Michael T. Hannan, and Guillem Pros. 2023. Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, 120(49):e2309350120.

Nadine Raaphorst. 2021. Administrative Justice in Street-Level Decision Making: Equal Treatment and Responsiveness. In *The Oxford Handbook of Administrative Justice*, pages 1–30. Oxford University Press.

Wassiliki Siskou and Ingrid Espinoza. 2024. "So, are you a different person today?" Analyzing Bias in Questions during Parole Hearings. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 116–128, Miami, Florida, USA. Association for Computational Linguistics.

Wassiliki Siskou, Laurin Friedrich, Steffen Eckhard, Ingrid Espinoza, and Annette Hautli-Janisz. 2022. Measuring plain language in public service encounters. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022) Potsdam, Germany*.

Tanya Stivers and N.J. Enfield. 2010. A coding scheme for question-response sequences in conversation. *Journal of Pragmatics*, 42(10):2620 – 2626. Question-Response Sequences in Conversation across Ten Languages.

Kalsoom Sumra. 2019. Social Equity in Public Administration: Fairness, Justice and Equity, tools for social change. *Pakistan Administrative Review*, 3(1):1–15.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Joanna Thornborrow. 2002. *Power talk : language and interaction in institutional discourse.* Real language series. Longman.

Bernardo Zacka. 2019. Street-Level Bureaucracy and Democratic Theory. In *Research Handbook on Street-Level Bureaucracy*, pages 448–461. Edward Elgar Publishing.

Paweł Łupkowski and Jonathan Ginzburg. 2017. Query responses. *Journal of Language Modelling*, 4(2):245–292.

## 8. Language Resource References

Ingrid Espinoza, Steffen Frenzel, Laurin Friedrich, Wassiliki Siskou, Steffen Eckhard, and Annette Hautli-Janisz. 2024. PSE v1.0: The First Open Access Corpus of Public Service Encounters. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13315–13320, Torino, Italia. ELRA and ICCL.

## A. Prompts

### A.1. Prompt for question classification

You are a linguistic expert specializing in question classification for German language.

Classify each question into exactly ONE of the following categories:

1. **Polar:** Yes/no questions seeking affirmation or negation
2. **Wh:** Information-seeking questions with interrogative words (*wer, was, wann, wo, warum, wie*, etc.)
3. **Alternative:** Questions presenting two or more explicit options connected by *oder*
4. **Tag:** Declarative statements with interrogative tags (*oder, ne, nicht*, etc.)
5. **Other:** Questions not fitting the above categories (echo, rhetorical, ambiguous)

Respond with ONLY the category name (*Polar, Wh, Alternative, Tag*, or *Other*). No explanation needed.

### A.2. Prompt for response classification

You are a linguistic expert specializing in conversation analysis and question-answer sequences in institutional dialogue.

You will be given a question-response pair from administrative encounters between government officials and citizens. Your task is to classify the response into exactly ONE of the following categories:

**Response Type Categories:**

IMPORTANT: Check the categories in order. Start with the most specific patterns first.

1. **No Response:** The question receives no verbal acknowledgment or reaction from the addressee. The turn is skipped or ignored. Look for empty responses, silence markers, or complete absence of a response.
   - Example: Q: "Haben Sie den Antrag?" A: "" (empty or "[silence]")
2. **Interrupted Response:** The response is cut off by another speaker BEFORE the answer is complete. The interruption marker ("/" or "//") must appear BEFORE the informational content is delivered or within the first sentence. If the first sentence provides a complete answer and interruption markers appear later in the utterance, it is NOT an interrupted response.
   - Example of Interrupted Response: Q: "Wo wohnen Sie?" A: "Ich wohne in Berl/"
   - Example of NOT Interrupted (complete answer despite later interruption): Q: "Haben Sie Kinder?" A: "Ja, zwei Kinder. Und meine Frau/"
3. **Response with Question:** The response takes the form of a question (counter-question), either seeking clarification before answering or challenging the premise of the original question. Look for question marks "?" in the response.
   - Example: Q: "Haben Sie den Antrag ausgefüllt?" A: "Welchen Antrag meinen Sie?"
4. **Non-Answer:** The response acknowledges the question but fails to provide the requested information. This includes evasive answers, hedges, topic shifts, or irrelevant responses. Common patterns include "Das weiß ich nicht", "Das ist schwierig", "Vielleicht", or changing the topic.
   - Example: Q: "Wann können Sie kommen?" A: "Das ist schwierig zu sagen."

5. **Proper Answer**: The response directly addresses the question with relevant and complete information. The answer fulfills the informational request. This should only be selected if none of the above categories apply.
   - Example: Q: "Haben Sie Kinder?" A: "Ja, zwei Kinder."
6. **Other**: Responses that don't fit the above categories, including ambiguous cases, minimal acknowledgments (backchannel signals like "mhm", "aha", "okay"), or responses combining multiple functions.
   - Example: Q: "Verstehen Sie?" A: "Mhm."

**Instructions**:

- Check each category in order from 1 to 6.
- Look for explicit markers: empty/silence for "No Response", "/" or "//" for "Interrupted Response", "?" for "Response with Question"
- Only classify as "Proper Answer" if the response clearly and directly answers the question
- Pay special attention to hedging language, topic shifts, and minimal responses
- Respond with ONLY the category name: "No Response", "Interrupted Response", "Response with Question", "Non-Answer", "Proper Answer", or "Other"
- Do not provide explanations or additional comments.

Respond with exactly one of these labels:

- No Response
- Interrupted Response
- Response with Question
- Non-Answer
- Proper Answer
- Other