# Identifying Small Talk in Natural Conversations

**Steffen Frenzel**[1] **and Annette Hautli-Janisz**[2]
[1]University of Potsdam, [2]University of Passau
steffen.frenzel@uni-potsdam.de, annette.hautli-janisz@uni-passau.de

## Abstract

Small talk is part and parcel of human interaction and is rather employed to communicate values and opinions than pure information. Despite small talk being an omnipresent phenomenon in spoken language, it is difficult to identify: Small talk is situated, i.e., for interpreting a string of words or discourse units, outside references such as the context of the interlocutors and their previous experiences have to be interpreted. In this paper, we present a dataset of natural conversation annotated with a theoretically well-motivated distillation of what constitutes small talk. This dataset comprises of verbatim transcribed public service encounters in German authorities and are the basis for empirical work in administrative policy on how the satisfaction of the citizen manifests itself in the communication with the authorities. We show that statistical models achieve comparable results to those of state-of-the-art LLMs.

## 1 Introduction

Small talk is an omnipresent phenomenon when people interact with each other. There is a variety of reasons why people engage in small talk, for instance to exhibit politeness, to build a connection with strangers or to start a conversation. From a linguistic point of view, small talk is a highly interesting type of conversation, for it is not primarily focused on the exchange of information – one could even argue that the topic of the conversation does not really matter – but rather about the exchange of values and opinions. From a computational point of view, small talk is a challenging phenomenon because it is highly context dependent, i.e., the individual background of the interlocutors together with the situational context determines the scope and content of the small talk. The genre they mostly appear in, namely conversations, is underrepresented in terms of resources overall, but in terms of small talk in particular.

But small talk is crucial for socio-linguistic analyses of conversations. The source of the data in this paper are public service encounters in Germany (Espinoza et al., 2024), i.e., direct conversations between citizens and representations of the state where citizens ask for support or benefits from the representatives. Previous work in administrative policy shows that even if the decision of the state is not in favor of the citizen, emphatic communication yields satisfaction scores that parallel those of favorable decisions (Guy et al., 2014). Therefore, being able to measure and identify relation-building blocks of conversation paves the way for meaningful sociolinguistic analyses of conversations at scale. The challenges are two-fold: From a theoretical point of view, concrete definitions for the concept of small talk are lacking, making the process of generating annotation guidelines tricky. Moreover, small talk is mostly performed in conversations – those are time-consuming to record and to transcribe, making sufficient training data expensive.

The contributions of this paper are three-fold: First, we put forward theoretically-motivated annotation guidelines that can be used to annotate small talk in transcribed conversations. We also present a new, human-annotated small talk dataset containing more than 2,600 utterances from German public service encounters. Lastly, we show that statistical models such as Logistic Regression or Support Vector Machines achieve results comparable to state-of-the-art LLMs after thorough training. Our error analysis demonstrates the difficulties of classifying small talk automatically.

## 2 Background

### 2.1 Theoretical conceptions of small talk

There is an abundance of literature on naming and defining the concept of small talk. It is investigated with a focus on its social functions (Fried-



*Proceedings of LaTeCH-CLfL 2025*, pages 272–277
May 4, 2025 ©2025 Association for Computational Linguistics

laender, 1922; Malinowski, 1949; Ventola, 1979; Coupland et al., 1992; Eggins and Slade, 2004; Senft, 2009; Chen et al., 2022), its impact on conversational structures (Laver, 1975; Edmondson and House, 1981; Schneider, 1988) and with respect to cultural differences (Isbister et al., 2000; Endrass et al., 2011). Regarding the topics covered in small talk, Schneider (1988) develops a taxonomy of topics distinguishing between topics concerning the immediate situation, the external situation and the communication situation. Isbister et al. (2000) shows that certain conversational topics are perceived as safe or unsafe depending on the cultural background of the subjects. In a follow-up study, Endrass et al. (2011) investigate how the prototypical distribution of conversation topics turns out for German and Japanese.

An example of what we consider small talk is shown in Figure 1. Prototypical topics according to Schneider's (1988) taxonomy appear ('family' and 'holidays'), but they are dependent on situational context (here, pre-christmas). These topics appear frequently in our dataset since they are connected to the main purpose of the conversation (applying for family benefits, for example). Other topics from Schneider's taxonomy (e.g., 'music' or 'sports') appear rarely or not at all. For the purpose of the annotation guidelines, we apply the theoretical concept of small talk topics to the conversational and cultural context of our dataset.

1. **Citizen:** Yes, in four weeks!
2. **Official:** Crazy, completely crazy!
3. **Citizen:** [laughs] And the children are already going crazy at home. I mean it's not normal anymore!
4. **Official:** Already? Because of Christmas?
5. **Citizen:** Yes, well I have decorated the house already, you know? So yes, they are really exited.
6. **Official:** Ah, nice!

Figure 1: Example of small talk (translated, German original transcript id: 202111240815el4d0y4nMAYMS)

### 2.2 Small talk in NLP

With the rise of conversational AI systems there has been a growing interest in modeling and generating small talk (also under the labels 'chitchat', 'informal conversation', 'off-topic' talk etc.) (Sun et al., 2021; Choudhary and Kawahara, 2022; Stricker and Paroubek, 2024b,a, inter alia). Different attempts were made to equip conversational agents with small talk functions (Bickmore and Cassell, 2001; Cavazza et al., 2010; Mattar and Wachsmuth, 2012; Zhao et al., 2022) since several studies indicate they can help establishing a personal bond with the user (Reeves and Nass, 1996; Morkes et al., 1998; Chao et al., 2021). Chiu et al. (2022) and Liu et al. (2023) focus on generating transitions from small talk to task-oriented dialogue.

For English, a few attempts to classify small talk have been made. Stewart et al. (2006) detect small talk in conversational telephone speech using supervised models, based on their taxonomy on simple lexical and syntactic features. Arguello and Rosé (2006) employ lexical and syntactic features into their classification model. Joty et al. (2013) develop an unsupervised topic segmentation model that detects small talk as 'off-topic' segments. Konigari et al. (2021) test for the first time a transformer-based model for off-topic detection in open-domain conversations. Lai et al. (2022) introduce a human-annotated dataset for chit-chat detection in English livestreaming videos.

For German, similar work is lacking. This carries over to studies using the latest generation of LLMs, which have not been tested on such a task and also not against traditional text classification models. This is the starting point of this paper: We introduce a novel dataset for German small talk[1] and show that statistical models are on a par with the latest generation of LLMs for predicting small talk in natural conversation.

## 3 Annotation study

### 3.1 Dataset

Our experiments are conducted on the PSE v1.0 dataset (Espinoza et al., 2024), a collection of verbatim transcribed Public Service Encounters in various German authorities that were recorded between 2021 and 2023. The dataset consists of 106 conversations with a total of more than 31,000 speaker turns and 433,780 tokens. PSEs are usually initiated by a citizen's application for social benefits. During those meetings the public official has to determine eligibility and extent of the support, which means that the conversations cover highly personal topics. The representatives are therefore interested in creating an open conversational atmo-

[1]The dataset and the full annotation guidelines are available on Github: https://github.com/steffrenzel/naacl_2025_smalltalk_detection

sphere, with small talk being one of the linguistic mechanisms to achieve this goal.

### 3.2 Manual annotation of small talk

For the scope of this paper, small talk is assumed to be polite conversation about light topics (Schneider, 1988). We refine the concept by having its purpose be the maintenance of social relations which are used to create a basis for the main discussion of a conversation. This kind of conversation is technically not restricted to certain topics, but it is usually about things that the speakers can easily agree on. Situational context and cultural background of the speakers can have an influence on the form of small talk, both on the length and the primary goal, as well as the choice of topics (Isbister et al., 2000; Mattar and Wachsmuth, 2012). We do not assume a constraint on the timing of small talk in conversations, because interlocutors can structure a conversation by continually inserting small talk sections (Schneider, 1988; Chen et al., 2022).

Based on these aspects we iteratively derive annotation guidelines by conducting manual multiple-person annotation rounds. Initial attempts with a 6-step Likert scale yield only slight agreement across annotators on individual speaker moves (on average 0.24 Cohen's Kappa). For the final dataset, we use complete conversations and subsequently annotate each speaker move with a binary value ('no small talk', 'small talk'), enabling the use of context in the prediction (more on this in Section 4). With this adjustment, agreement between the two annotators of the main study is 0.534 Cohen's Kappa for 700 speaker moves, which corresponds to moderate agreement (Viera and Garrett, 2005). Overall, Both annotators are native speakers of German and students in computational linguistics.

## 4 Predicting small talk

### 4.1 Training

We use four different models to identify small talk, two statistical models (Logistic Regression and SVM) and two language models (GBERT, GPT-4) to see how more expensive models fare in comparison with smaller models.

The baseline is Logistic Regression, with tf-idf vectorization for training and test set (German stopwords are removed by the vectorizer) and with sentence embeddings from `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019). For SVM, we again use tf-idf versus sentence embeddings and conduct 5-fold cross-validation with `StandardScaler` from `scikit-learn`. Fine-tuning is performed using `GridSearchCV`. Again, both tf-idf vectors and sentence embeddings are used for vectorization.

The first language model is GBERT, a BERT model specifically trained on German data (Chan et al., 2020), with the optimal settings determined by `GridSearchCV` (see Table 1). We also use GPT-4.o (OpenAI and et al., 2024) with zero-shot, few-shot and task framing prompting.

### 4.2 Results

The classification models show moderate performances. Interestingly, the final results of the different models are fairly close together, despite differing model complexity fine-tuning options. The full results are listed in Table 1. The weighted average F1-score is used here as the main evaluation metric.

LR needs thorough fine-tuning to achieve good results. Since we are dealing with an imbalanced dataset (the negative class is much more frequent than the positive class), the model tends to over-fit quickly and develops a bias to the majority class. To mitigate this, instead of the classes, the probabilities for each class are extracted and a manual decision boundary is applied to balance the output. This works fairly well and the final runs lead to the best overall results in the model comparison.

SVM performs slightly different in comparison to LR. In both cases, embeddings work significantly better than tf-idf vectors, which is to be expected. Despite the tf-idf vectors being less meaningful, SVM can still get reasonable results from them. In combination with sentence embeddings, the models performance is only slightly worse than the best run of LR.

The GBERT model leads to the worst overall performance. Training epochs and batch-size have to be kept small in order to mitigate over-fitting. The relatively small size of the training dataset in combination with the class imbalance again led to a biased classification. Several attempts were made to mitigate this effect, using class weights as well as minority class oversampling using SMOTE (Blagus and Lusa, 2013). However, these attempts did not lead to better performance.

Finally, we also test GPT-4.o using different prompting strategies. For the zero-shot runs, we just provide instructions but do not give any examples from our dataset, resulting in an F1-score

| Model | Vectorization | Hyperparameters | Acc | Prec | Rec | F1 | Support (0 / 1) |
|---|---|---|---|---|---|---|---|
| LR | tf-idf | `penalty=L2, solver=liblinear, boundary=0.16` | 0.51 | 0.75 | 0.51 | 0.56 | 514 (417/97) |
| | distilbert | `penalty=L2, solver=liblinear, boundary=0.2` | 0.71 | 0.75 | 0.71 | **0.73** | 514 (417/97) |
| SVM | tf-idf | `C=2.0, kernel='poly', gamma='auto', weight='balanced'` | 0.60 | 0.62 | 0.57 | 0.59 | 514 (417/97) |
| | distilbert | `C=.0, kernel='poly', gamma='auto', weight='balanced'` | 0.70 | 0.69 | 0.67 | 0.68 | 514 (417/97) |
| GBERT | - | `epochs=3, batch-size=16, warm-up-steps=500` | 0.61 | 0.66 | 0.53 | 0.59 | 514 (417/97) |
| GPT-4.o | - | `Few-Shot, temp=0.3` | 0.65 | 0.72 | 0.65 | 0.68 | 514 (417/97) |

Table 1: Best results across models and configurations, weighted average is used to account for class imbalance.

of 0.62. In the few-shot approach we add a few examples for both classes to the prompt. This approach works best, with an F1-score of 0.68. In the chain-of-thought run, we asks the model to explain its decisions, which does not work well since the model constantly predicts the negative class. For all these runs, the temperature is set to 0.3 – higher temperatures lead to less reproducible results and do not improve performance.

### 4.3 Error analysis

Both the manual annotation and the automatic classification show the difficulties in identifying small talk in our dataset. A qualitative analysis of the results shows major differences in how the classes are distributed over the course of a conversation.

Since the human annotators were given transcripts of complete conversations and their task was to classify on utterance level, they were aware of the conversational context. In both manual annotations, it is rare for a single utterance to be classified as small talk, while the surrounding utterances are not small talk. Instead, usually longer sections of a conversation are continuously identified as small talk - these occur particularly frequently at the beginning and end of a conversation. The biggest discrepancies between the two human annotators arise when identifying the transitions between small talk and other parts of the conversation. This shows once again that it is difficult to clearly distinguish small talk from other parts of conversation - there is often a 'transition zone' that can be interpreted differently despite comprehensive annotation guidelines.

Classification models that learn the concept of small talk only indirectly from the training data, on the other hand, often classify stand-alone utterances positively, while the surrounding utterances are classified negatively. Presumably, lexical and semantic criteria are more important here than the position in the conversation and the contextual utterances.

## 5 Conclusion

The error analysis has shown which problems remain in the classification of small talk. Complex classification models such as neural networks and transformer-based models are less suitable for this task until more training data is available. LLMs achieve in general good results in classifying the data, but prompting is the only way to control the classification. Simple classification models are labor-intensive as they have to be precisely fine-tuned. Ultimately, however, they provide the most transparent classifications and - at least in our study - achieved results comparable to those of LLMs.

## Limitations

Operationalizing the concept of small talk for this task remains the biggest challenge. We learned in the process of (re-)designing the manual annotation that conversational context is key information for the human annotators. However, this kind of

information needs to better implemented into the automatic classification, e.g. by engineering additional features.

## Acknowledgments

We would like to thank our student assistants Simon Bross, Jana-Linn Lauruschkus, Klymentii Myslyvyi and Anna-Kezia Rosenbauer for annotating the dataset and additional model testing. We would like to thank Diego Frassinelli for additional supervision and the reviewers for their helpful feedback.

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.

## References

Jaime Arguello and Carolyn Rosé. 2006. Topic-segmentation of dialogue. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 42–49, New York City, New York. Association for Computational Linguistics.

Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, page 396–403, New York, NY, USA. Association for Computing Machinery.

Rok Blagus and Lara Lusa. 2013. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14:106.

Marc Cavazza, Raul Santos de la Camara, and Markku Turunen. 2010. How was your day? a companion eca. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, page 1629–1630, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

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.

Chi Hsiang Chao, Xi Jie Hou, and Yu Ching Chiu. 2021. Improve chit-chat and QA sentence classification in user messages of dialogue system using dialogue act embedding. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 138–143, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Jiaxin Chen, Yudong Guo, and Jinyun Duan. 2022. How and when phatic communion enhances advice taking. *Asian Journal of Social Psychology*, 25(4):611–622.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158, Dublin, Ireland. Association for Computational Linguistics.

Ritvik Choudhary and Daisuke Kawahara. 2022. Grounding in social media: An approach to building a chit-chat dialogue model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 9–15, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Justine Coupland, Nikolas Coupland, and Jeffrey Robinson. 1992. "how are you?": Negotiating phatic communion. *Language in Society*, 21:207 – 230.

W. Edmondson and J. House. 1981. *Let's Talk, and Talk about it: A Pedagogic Interactional Grammar of English*. U-&-S-Pädagogik. Urban & Schwarzenberg.

S. Eggins and D. Slade. 2004. *Analysing Casual Conversation*. Equinox Textbooks and Surveys in Linguistics. University of Toronto Press.

Birgit Endrass, Yukiko Nakano, Afia Akhter Lipi, Matthias Rehm, and Elisabeth André. 2011. Culture-related topic selection in small talk conversations across germany and japan. In *Intelligent Virtual Agents*, pages 1–13, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Violet Helen Friedlaender. 1922. *Pied Piper's Street and Other Essays*. Arrowsmith.

Mary Guy, Meredith Newman, and Sharon Mastracci. 2014. *Emotional Labor. Putting the Service in Public Service*. Routledge.

Katherine Isbister, Hideyuki Nakanishi, Toru Ishida, and Cliff Nass. 2000. Helper agent: designing an assistant for human-human interaction in a virtual meeting space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, page 57–64, New York, NY, USA. Association for Computing Machinery.

S. Joty, G. Carenini, and R. T. Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.

Rachna Konigari, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic shift detection for mixed initiative response. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 161–166, Singapore and Online. Association for Computational Linguistics.

Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. 2022. BehanceCC: A ChitChat detection dataset for livestreaming video transcripts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7284–7290, Marseille, France. European Language Resources Association.

John Laver. 1975. Communicative functions of phatic communion. In Adam Kendon, Richard M. Harris, and Mary R. Key, editors, *Organization of Behavior in Face-to-Face Interaction*, pages 215–238. De Gruyter Mouton, Berlin, New York.

Ye Liu, Stefan Ultes, Wolfgang Minker, and Wolfgang Maier. 2023. System-initiated transitions from chitchat to task-oriented dialogues with transition info extractor and transition sentence generator. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 279–292, Prague, Czechia. Association for Computational Linguistics.

Bronislaw Malinowski. 1949. The problem of meaning in primitive languages. In *The meaning of meaning*. Routledge & Kegan Paul.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat. In *KI 2012: Advances in Artificial Intelligence*, pages 119–130, Berlin, Heidelberg. Springer Berlin Heidelberg.

John Morkes, Hadyn K. Kernal, and Clifford Nass. 1998. Humor in task-oriented computer-mediated communication and human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, CHI '98, page 215–216, New York, NY, USA. Association for Computing Machinery.

OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people and pla. *Bibliovault OAI Repository, the University of Chicago Press*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

K.P. Schneider. 1988. *Small Talk: Analyzing Phatic Discourse*. Linguistic (Hitzeroth). Hitzeroth.

Gunter Senft. 2009. Phatic communion. In *Culture and language use*. John Benjamin.

Robin Stewart, Andrea Danyluk, and Yang Liu. 2006. Off-topic detection in conversational telephone speech. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 8–14, New York City, New York. Association for Computational Linguistics.

Armand Stricker and Patrick Paroubek. 2024a. Chitchat as interference: Adding user backstories to task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3203–3214, Torino, Italia. ELRA and ICCL.

Armand Stricker and Patrick Paroubek. 2024b. A few-shot approach to task-oriented dialogue enhanced with chitchat. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 590–602, Kyoto, Japan. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.

Eija Ventola. 1979. The structure of casual conversation in english. *Journal of Pragmatics*, 3.

Anthony Viera and Joanne Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family medicine*, 37:360–3.

Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. 2022. UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland. Association for Computational Linguistics.