# Automatic Identification of Generalizing Passages in German Fictional Texts using BERT with Monolingual and Multilingual Training Data

**Thorben Schomacker**
Hamburg Univ. of Applied Sciences
thorben.schomacker@haw-hamburg.de

**Tillmann Dönicke**
Univ. of Göttingen
tillmann.doenicke@uni-goettingen.de

**Marina Tropmann-Frick**
Hamburg Univ. of Applied Sciences
marina.tropmann-frick@haw-hamburg.de

## 1 Introduction

This paper is concerned with the automatic identification of generalizing passages like *all ducks lay eggs* or *tigers are usually striped* (cf. Leslie and Lerner, 2016). In fictional texts, these passages often express some sort of (self-)reflection of a character or narrator or a universal truth which holds across the context of the fictional world (cf. Lahn and Meister, 2016, p. 184), and therefore they are of particular interest for narrative understanding in the computational literary studies.

In the following, we first establish a new state of the art for detecting generalizing passages in German fictional literature using a BERT model (Devlin et al., 2019). In a second step, we test whether the performance can be further improved by adding samples from a non-German corpus to the training data.

## 2 Data and Previous Work

The Modes of Narration and Attribution Corpus (**MONACO**) (Barth et al., 2021) is a corpus of German fictional texts from 1600 to 1950. It is annotated with three basic narratological phenomena, each having its own tagset. The annotations are performed on the supra-clause level, meaning that an annotated passage consists of at least one and possibly multiple subsequent clauses. In this paper we only consider the phenomenon *Generalizing Interpretation* (GI). GI passages are defined as quantified statements about entities, situations, locations etc. which are at least partially unknown to the speaker. GI is annotated with six tags, which represent various types of quantification (Dönicke et al., 2021): ALL (universal), MEIST (majority), EXIST (existential), DIV (vague), BARE (generic), NEG (negated). As of version 3.0, MONACO consists of 13,375 annotated clauses, from which 18.3% are part of a GI passage. The task of identifying generalizing passages in MONACO was previously approached by Gödeke et al. (to appear, cf. Varachkina et al. (2022)). The authors trained a clause-level statistical classifier and we use their method as baseline.

In addition to MONACO, we use the multi-genre corpus constructed within the Situation Entities (**SITENT**) project[1], which contains annotations for five situation entity types (Friedrich, 2018), also on clause level. One of these entity types are *General Statives* (Friedrich et al., 2015, p. 30; Smith, 2003, p. 24), which describe regularities of events (subclassified as *generalizing sentences*) or properties of kinds of entities (subclassified as *generic sentences*) and are therefore very similar to the GI passages in MONACO. The corpus consists of 50,009 annotated clauses, from which 18.1% are part of a general stative.

## 3 Method

To harmonize the annotations in MONACO and SITENT, we relabeled the clauses annotated as general stative in SITENT with the GI tags of MONACO using a list of quantifier lexemes for each tag (e.g. *every* $\mapsto$ ALL) and a small set of rules to connect clauses to passages. For all experiments we use two texts from MONACO each as development set (Gellert, Fontane) and test set (Wieland, Seghers). Furthermore, we use three different training sets: 1) We use the MONACO training set, consisting of all other texts in MONACO. 2) We merge the MONACO training set with the complete SITENT corpus to create the novel dataset CAGE ("Clauses Annotated with Generalization Phenomena"). CAGE consists of about 21% clauses from MONACO and 79% clauses from SITENT. 3) We create CAGE-small which only includes those texts from SITENT that are categorized as fiction, essays, ficlets and letter. This should lead to a more homogeneous dataset in terms of writing style and an

[1] https://github.com/annefried/sitent/tree/master/annotated_corpus

| Training Set | Method | Binary | Multi | ALL | BARE | DIV | EXIST | MEIST | NEG |
|---|---|---|---|---|---|---|---|---|---|
| MONACO | RandomForest | – | 0.27 | 0.39 | 0.18 | 0.44 | 0.40 | 0.00 | 0.19 |
| MONACO | G-BERT | **0.78** | **0.61** | 0.51 | **0.50** | 0.50 | **0.60** | **1.00** | **0.57** |
| CAGE | G-BERT | 0.76 | 0.54 | **0.57** | **0.50** | 0.39 | 0.57 | 0.75 | 0.46 |
| CAGE-small | G-BERT | 0.76 | 0.54 | 0.54 | 0.44 | **0.54** | 0.33 | **1.00** | 0.37 |

Table 1: Performance of Gödeke et al.'s random forest and our BERT models on the test data, showing clause-level F1 for the binary task as well as macro-averaged F1 and class-wise F1s for the multi-label task.

equal distribution of both languages. CAGE-small consists of about $51\%$ clauses from MONACO and $49\%$ clauses from SITENT.

We use the data for two tasks: 1) **binary** classification, i.e. whether a clause is part of a generalizing passage or not, and 2) **multi**-class classification with the six sub-tags. Since the tags are not exclusive, the latter task is also a multi-label task.

**Sample Format** Since clauses are the minimal annotation unit, we use contextualized clauses as classification input, which means that the input text for one clause consists of the current clause's sentence and the two neighboring sentences. To mark the current clause, we insert HTML-style ** tags around it. In addition to the six GI tags, we use the label NONE for our experiments, which is always assigned when a clause has no original label, to make two things possible: 1) to calculate special loss-weights (as described below) and 2) to have separate XAI reports for the features that lead to a classification where no original label is assigned.

**Hyperparameter Optimization** We use a BERT model and a batch size of 8 in all experiments, and trained the model for 20 epochs.[2] We consecutively optimized the model parameters on the development set in each experiment, starting with the pre-trained model. We tested *bert-german-base-cased* (Chan et al., 2019), *bert-base-cased* (Devlin et al., 2019), *bert-multilingual-base-cased* (Devlin et al., 2019), *roberta-base-wechsel-german* (Minixhofer et al., 2022) and *gbert-large* (Chan et al., 2020), where *gbert-large* outperformed the others by far. For the optimizer, we compared LAMB (You et al., 2020) and ADAM (Kingma and Ba, 2017) with the learning rate $lr \in \{1e-3, 1e-4, 1e-5, 1e-6\}$. LAMB with $1e-4$ performed the best. Similarly to El Anigri et al. (2021), we use hidden dropout and attention dropout and optimize the dropout probability $p \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. Hidden dropout with 0.3 and attention dropout with 0.0 performed best. For the multi-label classification models, we experimented with adding weights to positive samples for each class in the loss function (so-called *pos weights*). We compared no weights, neg-scaled weights and none-scaled weights. For neg-scaled weights, the pos weight is $w_i = n_i/p_i$, where $n_i$ is the number of negative samples and $p_i$ the number of positive samples for the label $i$. For none-scaled weights, the pos weight is $w_j = (p_n + n_n)/n_n$, where $p_n$ is the number of positive samples and $n_n$ the number of negative samples for NONE and $j$ covers all labels but NONE. We found that the performance slightly declines when none- and neg-weights were applied, so we did not use loss weights in all experiments.

## 4 Results and Discussion

Table 1 shows the results on the test set. Since Gödeke et al. trained their model on MONACO version 1.1 with fewer texts, we retrained their model on the data split which we used and use its performance as baseline for the multi-label task. Our models outperform this baseline and further achieve relatively high results in the binary task.[3] Interestingly, adding the English data did not improve the models' overall performance.[4]

Explainable methods in artificial intelligence (XAI) help to add a more qualitative-level angle to the evaluation of the model. We used a combination of Local Interpretable Model-Agnostic Explanations (Lime) (Ribeiro et al., 2016) for feature importances and anchors (Ribeiro et al., 2018) to determine classification key words. One example is included in the appendix A and more examples are shown on the poster.

[2]Early stopping did not improve the performance.

[3]For comparison: Friedrich et al. (2016) achieve 29% F1 for generalizing sentences and 68% F1 for generic sentences when training and testing on the SITENT corpus.

[4]Our models are available at https://github.com/tschomacker/generalizing-passages-identification-bert.

## References


Florian Barth, Tillmann Dönicke, Benjamin Gittel, Luisa Gödeke, Anna Mareike Hofmann, Anke Holler, Caroline Sporleder, and Hanna Varachkina. 2021. MONACO: Modes of Narration and Attribution Corpus.

Branden Chan, Timo Möller, Malte Pietsch, and Tanay Soni. 2019. German BERT | State of the Art Language Model for German NLP.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. Technical Report arXiv:2010.10906, arXiv. ArXiv:2010.10906 [cs] type: article.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tillmann Dönicke, Luisa Gödeke, and Hanna Varachkina. 2021. Annotating quantified phenomena in complex sentence structures using the example of generalising statements in literary texts. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 20–32, Groningen, The Netherlands (online). Association for Computational Linguistics.

Salma El Anigri, Mohammed Majid Himmi, and Abdelhak Mahmoudi. 2021. How BERT's Dropout Fine-Tuning Affects Text Classification? In *Business Intelligence*, pages 130–139, Cham. Springer International Publishing.

Theodor Fontane. 2012. Der Stechlin. In *TextGrid Repository*. Digitale Bibliothek.

Annemarie Friedrich. 2018. Situation Entities corpus.

Annemarie Friedrich, Kleio-Isidora Mavridou, and Alexis Palmer. 2015. Situation entity types (annotation manual). Version 1.1.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

Christian Fürchtegott Gellert. 2012. Das Leben der schwedischen Gräfin von G... In *TextGrid Repository*. Digitale Bibliothek.

Luisa Gödeke, Florian Barth, Tillmann Dönicke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler, and Caroline Sporleder. to appear. Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung. *Zeitschrift für digitale Geisteswissenschaften*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Silke Lahn and Jan Christoph Meister. 2016. *Einführung in die Erzähltextanalyse*, 3 edition. J.B. Metzler, Stuttgart.

Sarah-Jane Leslie and Adam Lerner. 2016. Generic Generalizations. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2016 edition. Metaphysics Research Lab, Stanford University.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. Technical Report arXiv:2112.06598, arXiv. ArXiv:2112.06598 [cs] type: article.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Number: 1.

Anna Seghers. 2013. *Das siebte Kreuz*. Aufbau Verlage GmbH.

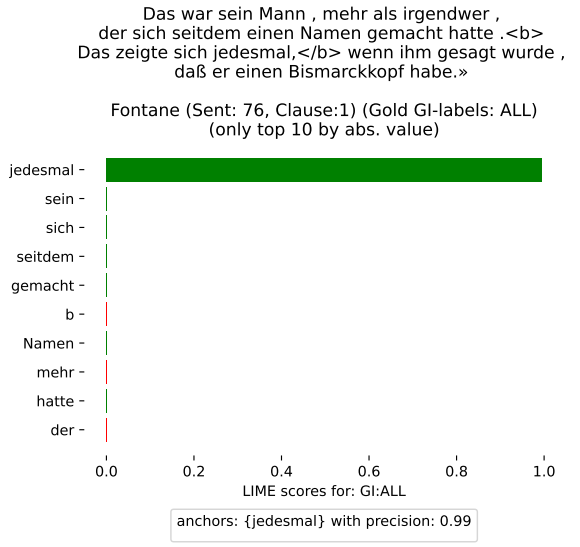
Carlota S. Smith. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge University Press. Google-Books-ID: okPPSq1G1OQC.

Hanna Varachkina, Florian Barth, Luisa Gödeke, Anna Mareike Weimer, and Tillmann Dönicke. 2022. Reflexive Passagen und ihre Attribution. A poster presentation at the 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" - DHd 2022 Kulturen des digitalen Gedächtnisses.

Christoph Martin Wieland. 2012. Geschichte des Agathon. In *TextGrid Repository*. Digitale Bibliothek.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. *arXiv:1904.00962 [cs, stat]*. ArXiv: 1904.00962.

# A Appendix




This was his man, more so than anyone who had made a name for himself since. **This showed up every time** he was told he had a Bismarck's head.

Figure 1: Lime scores and anchor(s) for ALL in Fontane (2012) on the cage-small model