

HeidelGram

Crossing the Boundary of Time: Fine-Tuning Modern NLP Models for Specialized Historical Corpus Data

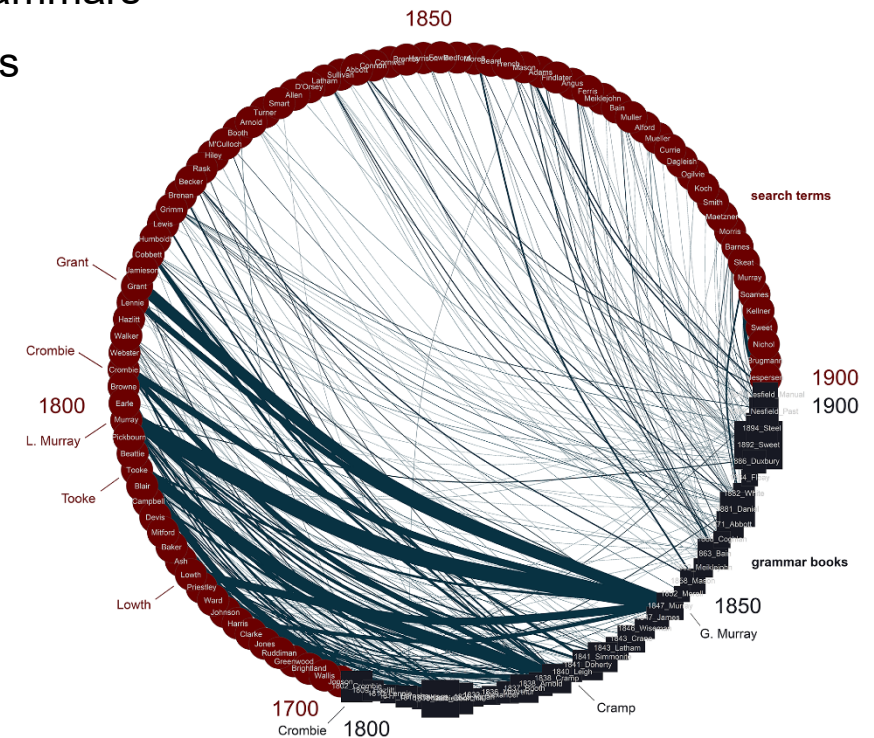
1. HeidelGram – Project Overview

Aims

- Compile and analyze a corpus of 16th-19th-century historical English grammars
- Innovatively combine corpus linguistic and network analytical approaches

Previous Work

- Compilation and analysis of a 19th-century pilot corpus (Busse et al., 2018; 2019)
- Grammarians' reference categories (based on 19th-century data) (Busse et al., 2020)
- Building a network of 16th-century grammarians and references (Busse et al., 2021)



Funded by



Deutsche
Forschungsgemeinschaft
German Research Foundation

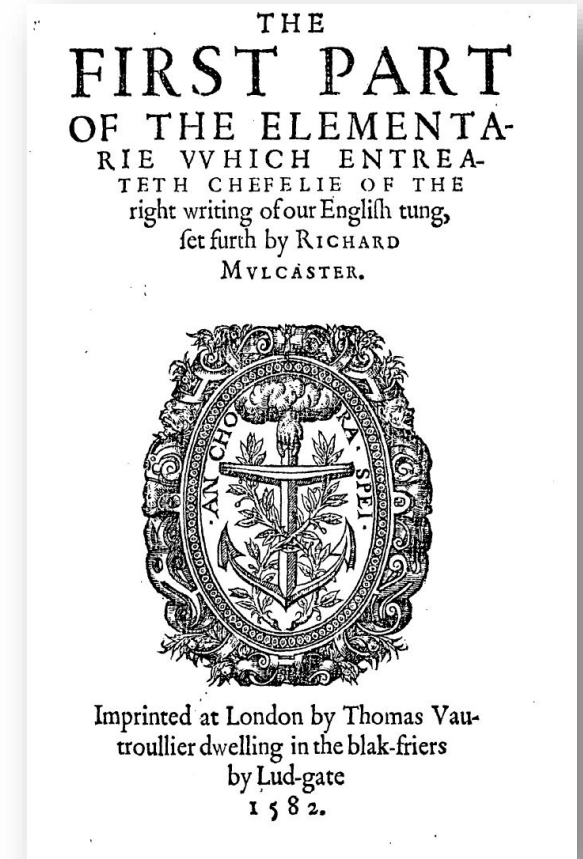


1. HeidelGram – Project Overview

Corpus of 16th-Century British Grammars

Conventional approaches to defining grammar (see McCarthy, 2020) combined with **verbal hygiene** (Cameron, [1995] 2012).

- I. Bullokar, William (1586) *Brief Grammar for English* (17,606 words)
- II. Coote, Edmund (1596) *The English Schoole-Maister Teaching all his Schollers* (29,476 words)
- III. Meurier, Gabriel (1586) *The Coniugations in Englishe and Netherdutche* (7,131 words)
- IV. Mulcaster, Richard (1582) *The First Part of the Elementarie Which Entreateth Chefelie of the Right Writing of our English Tung* (10,1047 words)
- V. Sherry, Richard (1577) *A Treatise of the Figures of Grammer and Rhetorike* (27,368 words)



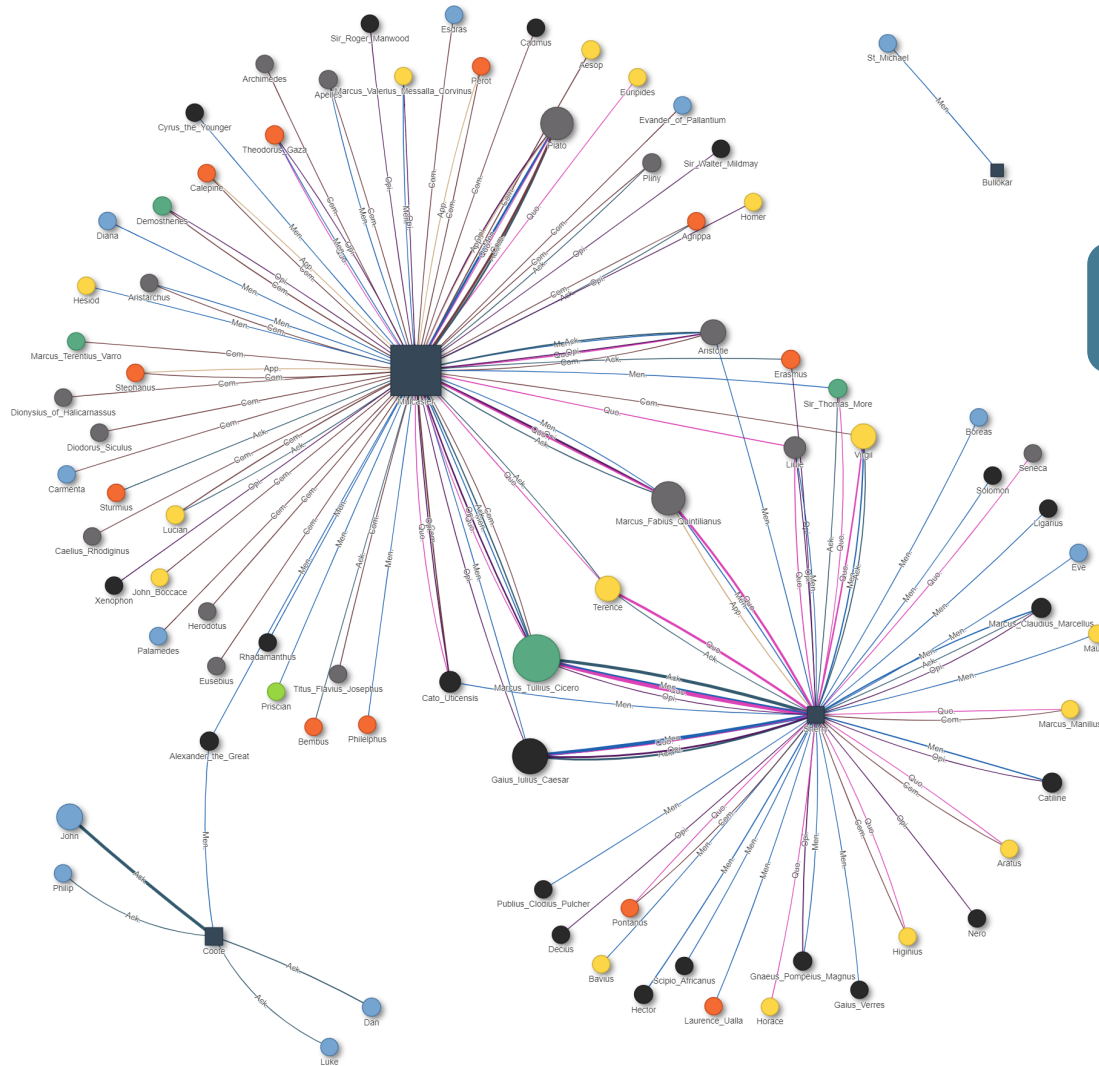
Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation



1. HeidelGram – Project Overview



<https://heidelgram.de/conference/cl2021/network.html>



“But why didn’t you use NER for this?” – CL2021

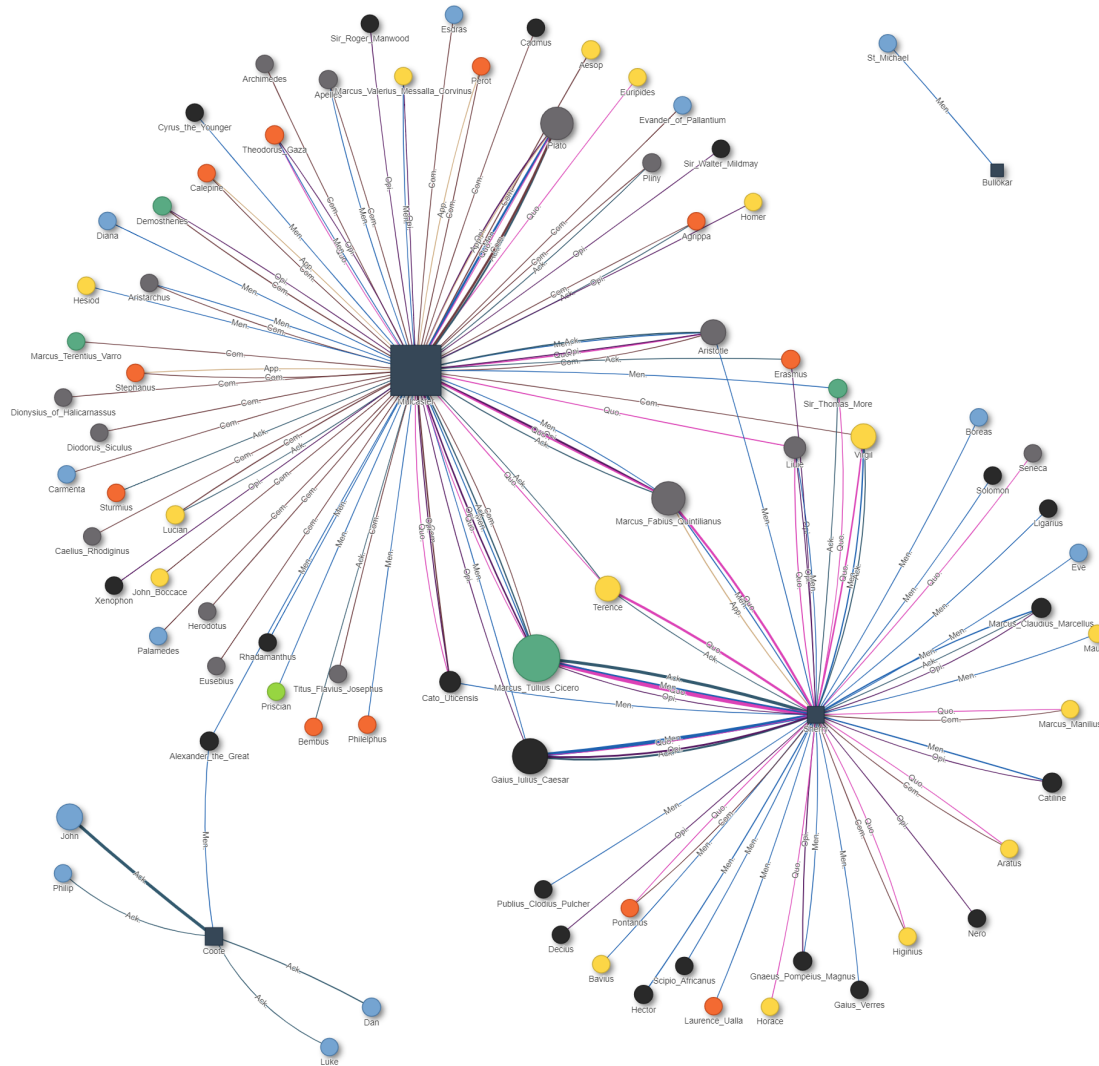
Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation



1. HeidelGram – Project Overview



<https://heidelgram.de/conference/s/cl2021/network.html>



Funded by

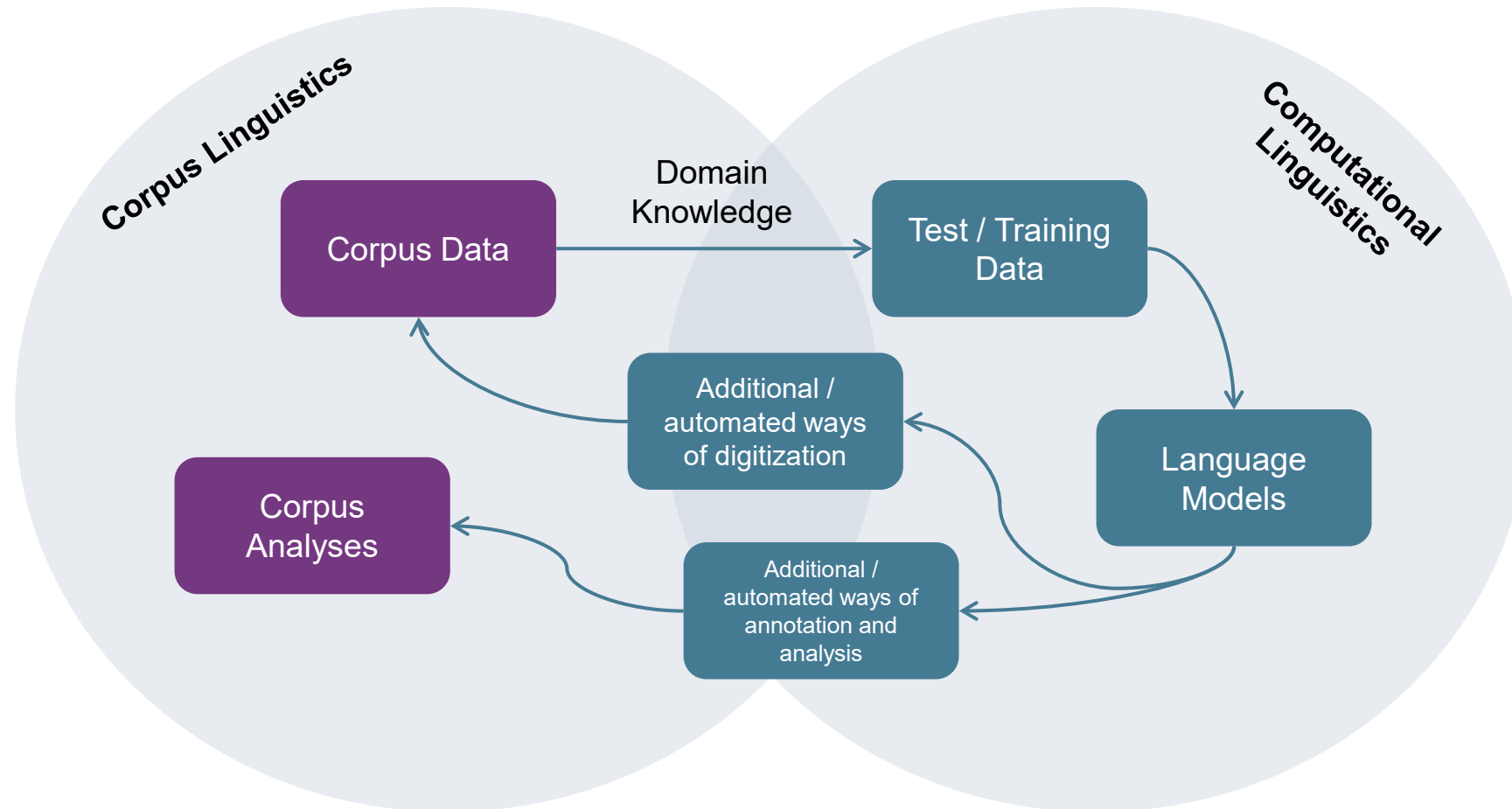


Deutsche
Forschungsgemeinschaft
German Research Foundation



1. HeidelGram – Project Overview

The Link Between Corpus and Computational Linguistics (NLP)



Funded by

DFG

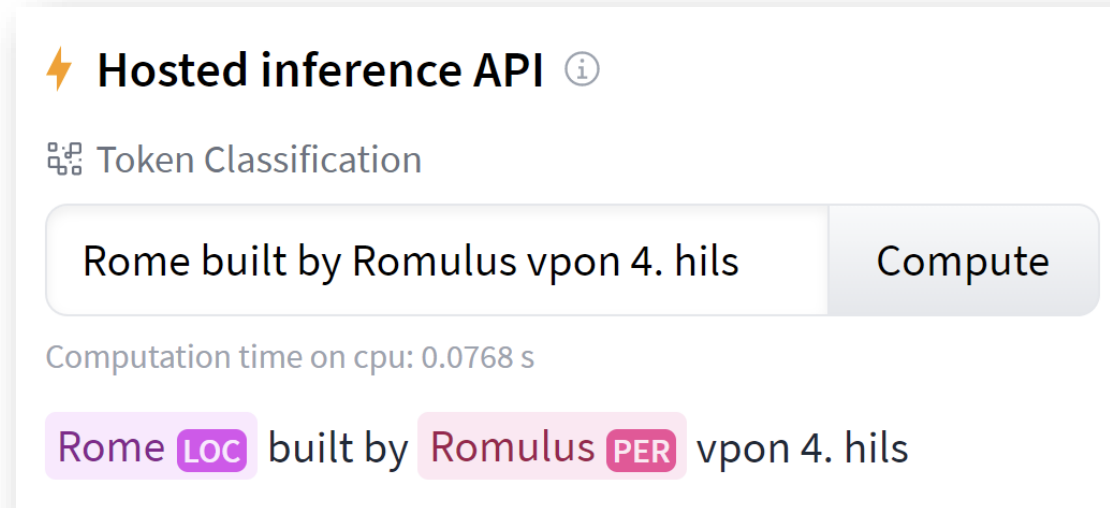
Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



“The task of **named entity recognition (NER)** is to find spans of text that constitute proper names and tag the type of the entity.”

(Jurafsky & Martin, 2020, p. 153)



The screenshot shows a web interface for a hosted inference API. At the top, there is a lightning bolt icon followed by the text "Hosted inference API" and an information icon. Below this, the text "Token Classification" is displayed with a small icon. A text input field contains the sentence "Rome built by Romulus vpon 4. hils", and a "Compute" button is to its right. Below the input field, the text "Computation time on cpu: 0.0768 s" is shown. The output of the inference is displayed below, with "Rome" highlighted in a purple box and labeled "LOC", "built by" in the original text, "Romulus" highlighted in a pink box and labeled "PER", and "vpon 4. hils" in the original text.

[huggingface.co / Model: dslim/bert-base-NER](https://huggingface.co/dslim/bert-base-NER)

2. Aims of this Paper

- Evaluate whether contemporary models can be fruitfully fine-tuned using historical language data
- Fine-tune modern transformer-based language models to perform NER on historical data, specifically 16th-century data (e.g., Schweter & Baiter, 2019)
- Long-term solution for multiple applications within the HeidelGram project (e.g., PoS tagging, OCR correction, text classification)
- Explore additional ways of analyzing (large amounts of) historical data using state-of-the-art NLP approaches



Disclaimer: The findings shown are preliminary and the models are under active development.
(Pilot Study)

Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation



3. Theoretical Background

How are we doing this?

- **Transfer learning**: transmitting as much knowledge as possible from the source setting to the target task or domain (Ruder, 2019, p.44).
- **Fine-tuned language model** = specialized to a domain (e.g., historical data) and/or downstream tasks (e.g., NER)

NLP Tasks

Masked Language Modeling (MLM)

A model which produces predictions for a masked token based on all left and right context (in the case of BERT)

(Devlin et al., 2019)

Named Entity Recognition (NER)

“The task of **named entity recognition (NER)** is to find spans of text that constitute proper names and tag the type of the entity.”

(Jurafsky & Martin, 2020, p. 153)

An equiuc is
a word ha-
uing diuers
meanings,
yet of one
part of spech;

Masked Language Modeling (MLM)

An equiuc is a **[MASK]** hauing diuers meanings, yet of one part of spech. (word)

(Bullokar, 1586, p.16)

Model	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5
HistBERT-16	word	letter	tongue	thing	Language
HeidelBERT-16	thing	verb	word	place	number
HeidelHistBERT-16	word	thing	letter	tongue	writing

The language models learn by performing/solving this task.

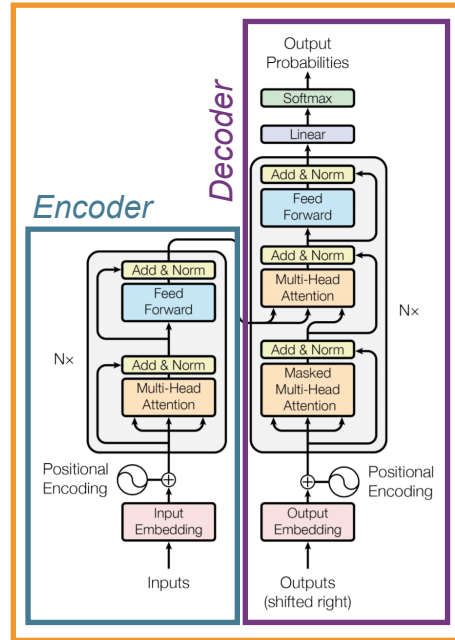
3. Theoretical Background

Natural Language Processing (NLP) is “the set of methods for making human language accessible to computers” (Eisenstein, 2019, p. 1) and ultimately “the ability for a computer/system to truly understand human language and process it in the same way that a human does” (Goyal et al., 2018, p. 16).

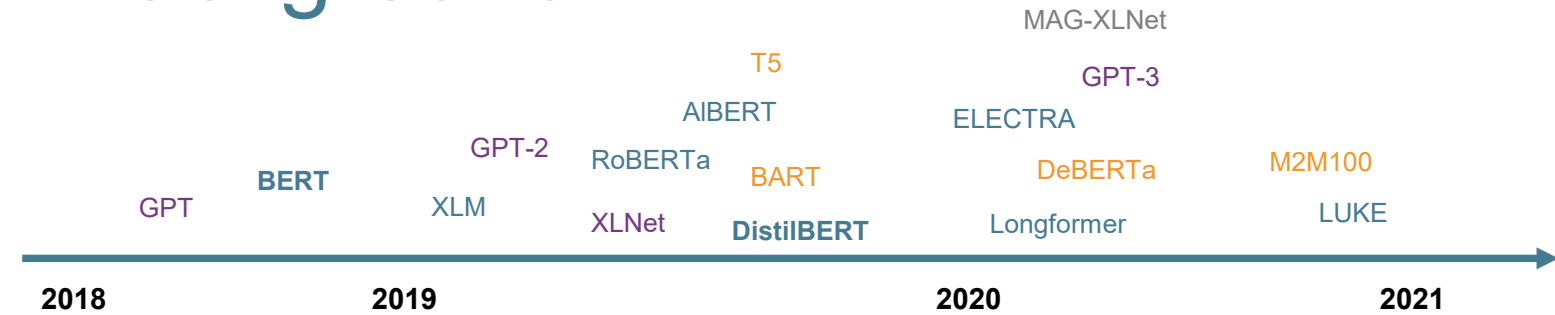
“Models that assign probabilities to sequences of words are called **language models** or **LMs**” (Jurafsky & Martin, 2020, p. 30).

Transformers are state-of-the-art deep learning models used in NLP. Their “key innovation [lies in] the use of self-attention layers.” (Jurafsky & Martin, 2020, p. 191).

3. Theoretical Background



“Attention is All You Need”
(Vaswani et al., 2017)



Auto-Encoding Transformer Models (best used for tasks **relying on the understanding of the input**)

→ e.g., token/document classification, sentiment analysis, ...

Auto-Regressive Transformer Models (best used for tasks that **generate language**)

→ e.g., text generation

Sequence-to-Sequence Transformer Models (best used for **input-dependent generative tasks**)

→ e.g., translation, summarization

Multimodal Transformer Models (best used for understanding **multimodal texts**) (*early-stage research*)

→ e.g., multimodal sentiment analysis

Funded by



Deutsche
Forschungsgemeinschaft
German Research Foundation



3. Theoretical Background



BERT (Bidirectional Encoder Representations from Transformers)

Developed by/at Google (Devlin et al., 2019; Zöllner et al., 2021)

- Non-traditional LM: BERT does not predict word/sentence probability! → *Masked Language Modeling & Next Sentence Prediction*
- Bidirectional (LR, RL) auto-encoding transformer model with multi-headed self-attention and 12 (24) stacked encoders
- Originally trained on *English Wikipedia* (2,500 million words) and *BooksCorpus* (800 million words)

We focus on **BERT/DistilBERT**

- BERT is an older, relatively well-understood model/architecture that is known to perform well on token classification tasks.
- It is flexible in terms of picking up on domains (Wei et al., 2021) and outperforms other models in terms of precision, recall, and F1-score (Mozafari et al., 2019).

Specifically, we use **DistilBERT** (Sanh et al., 2019)

- DistilBERT is a smaller, faster, cheaper, and lighter alternative to BERT (*bert-base-uncased*) based on distilling (teacher-student learning) → Faster and cheaper training (and inference) on cheap/consumer hardware
- It retains 97% of BERT's original performance (*GLUE*)
- In the future, we also want to have a look at RoBERTa which has similar benefits

Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation



4. Data

Fine-Tuning Data (Corpora)

- 16th-century component of EEBO ([EEBO-TCP](#), Phases 1 & 2) (5,210 documents, approx. 190 million tokens)
- 16th-century component of the HeidelGram corpus (5 grammars, approx. 182,628 tokens)

Input Data (Model)

- Layers of tokenization
 1. Word-based tokens with NER tags
 2. Sub-word tokenization needed for (Distil)BERT (to avoid OOV words)
 3. De-tokenization: recreating words out of sub-words
- Fixed-length sequences of tokens
- Simplified (IOB) tags for NER

NER (task)

William Bullokar to the
Raedor.

William I-PER
Bullokar B-PER
to 0
the 0
Raedor 0
. 0

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln

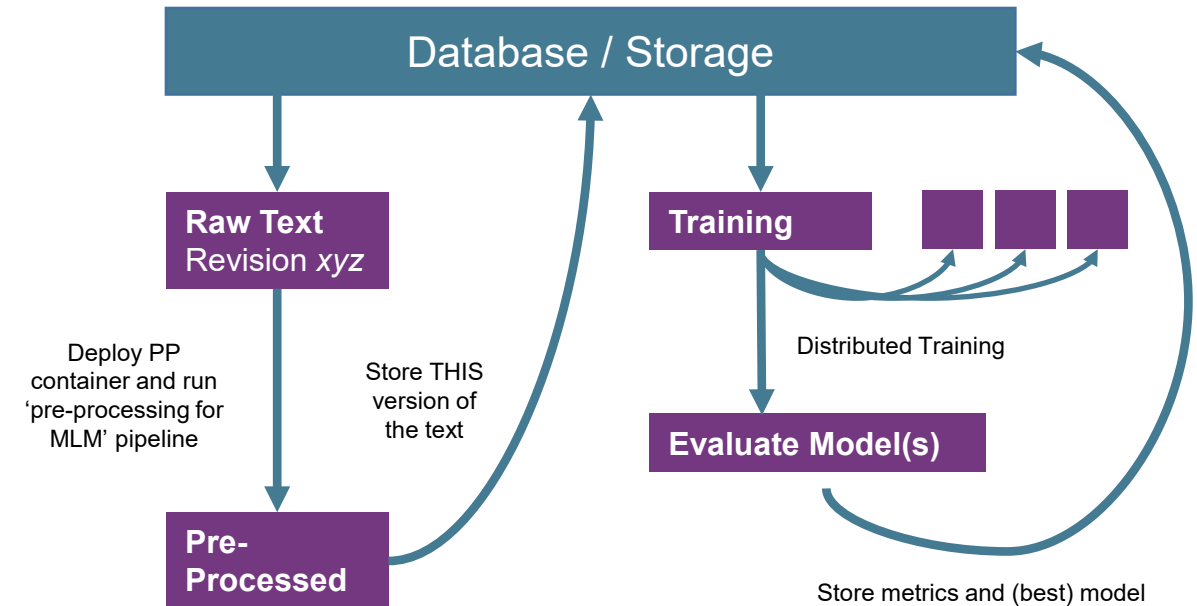


4. Data

We are currently moving HeidelGram to a **MySQL** and **code-based** (*Everything-as-Code*) architecture.

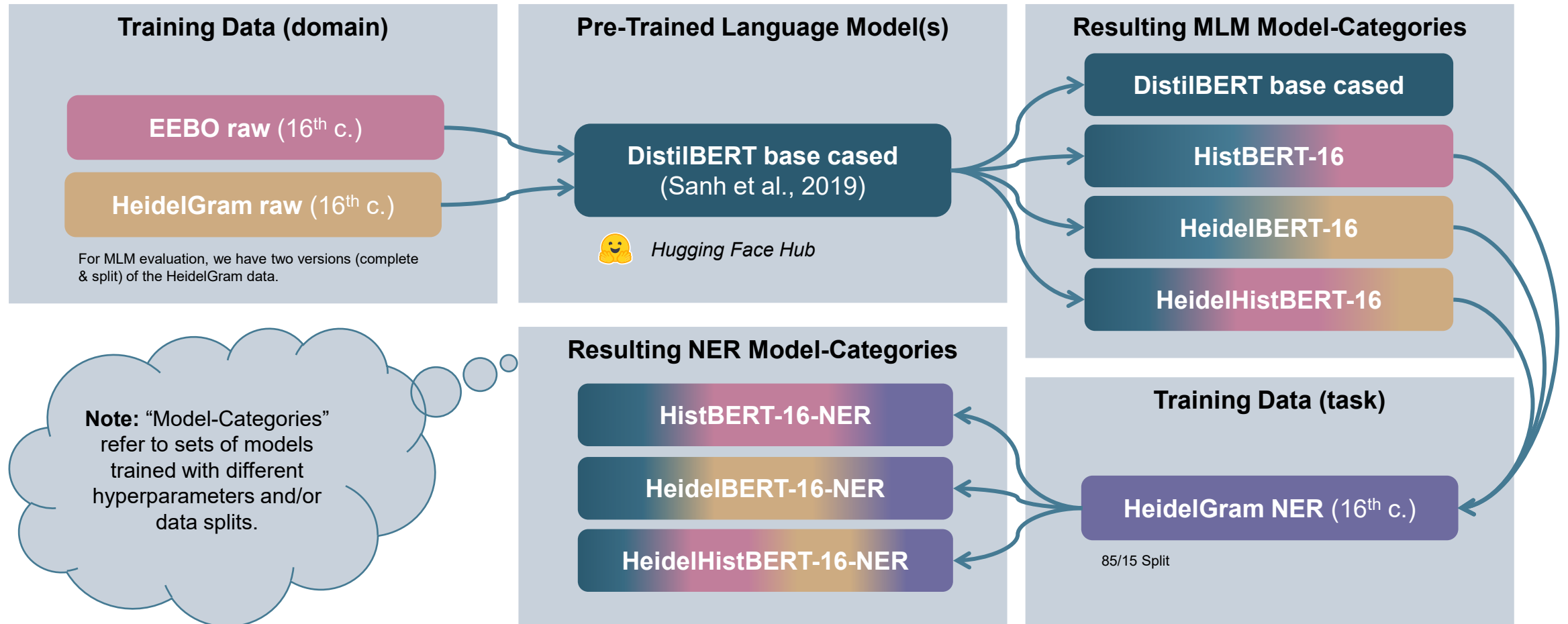
- *MySQL* Database as a shared **single source of truth** (SSOT) + *Gitlab* for code/config and as a container registry + file storage
- Various **benefits** such as ...
 - Relational layers of abstraction
(*grammar > editions > version(s) of the text > ...*)
 - Relating linguistic/textual data with extra-linguistic data
(*e.g., information about authors; metadata*)
 - Multiple layers of annotation (stand-off annotation)
 - Standardized, reproducible, shareable, and scalable data processing and research pipelines
 - Relative ease of performing (parts of) the training, analysis, etc. in the cloud

Pipeline Example: Automated Distributed Training

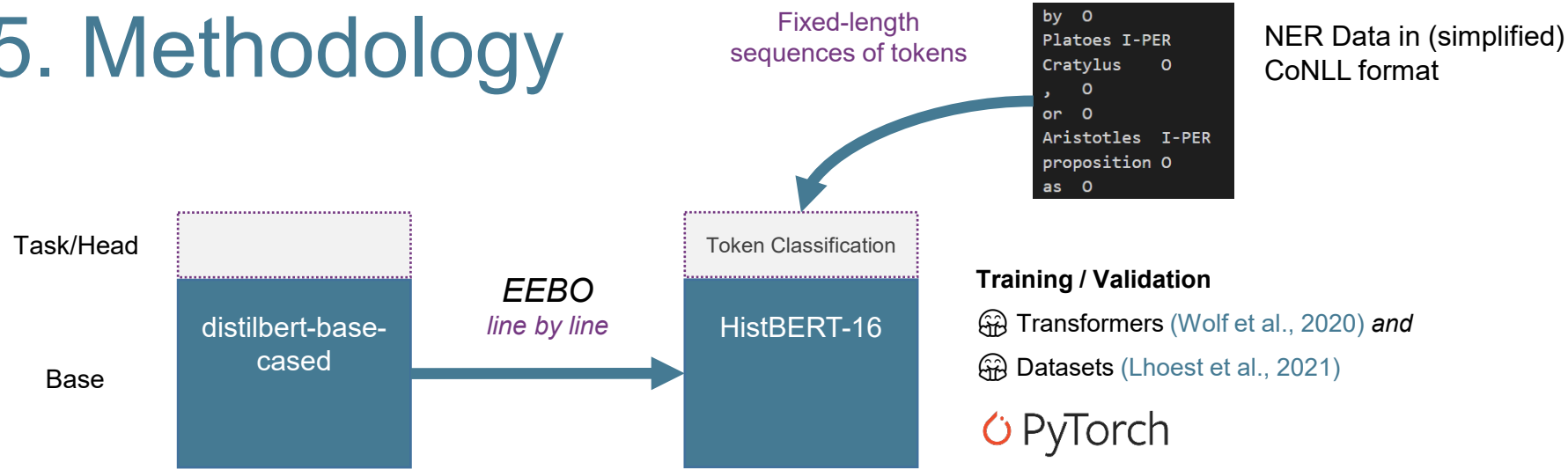


This whole process is defined via code (and/or configuration) and runs without human intervention. Also: *Self-Documenting*

5. Methodology



5. Methodology



During **fine-tuning**, the last layers (head) of the model are replaced based on what the model is supposed to do (e.g., token classification).

- Various 85-15 training/validation splits
- Our data **pre-processing** is still very crude (*Future Work*).
- Splitting up the 16th-century grammars into, for example, sentences would be helpful for training but poses challenges.
- For historical data it might be wise to train a new subword tokenizer as the set of 'common words', which are not split up, has changed over time.

Everything is based on **subword tokenization** (*WordPiece* for DistilBERT)

```
1 tokens = tokenizer('The Heidelberg project')
2 [tokenizer.convert_ids_to_tokens(id) for id in tokens['input_ids']]

['[CLS]', 'The', 'He', '##ide', '##l', '##G', '##ram', 'project', '[SEP]']
```

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

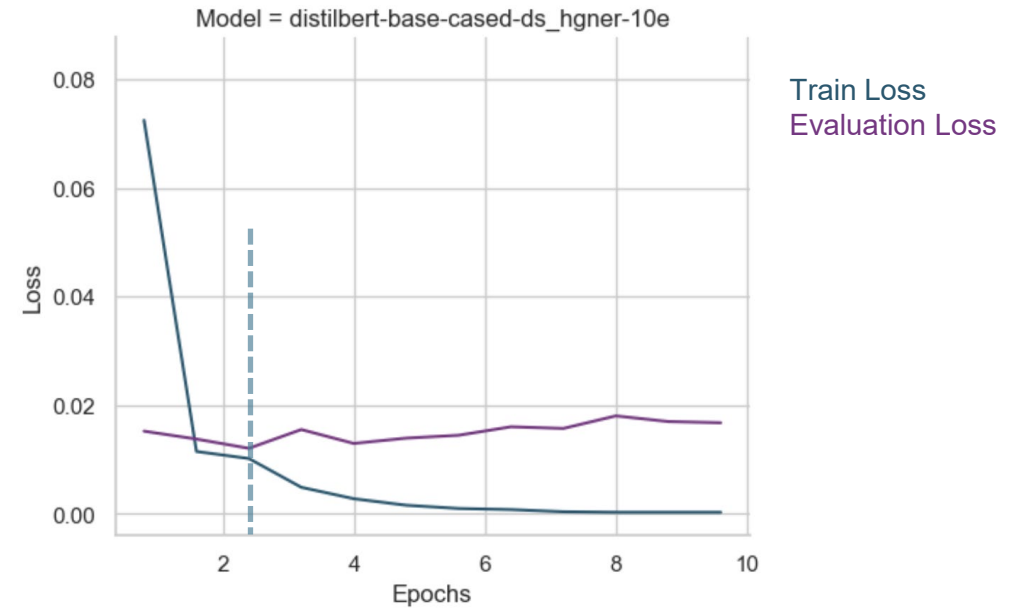
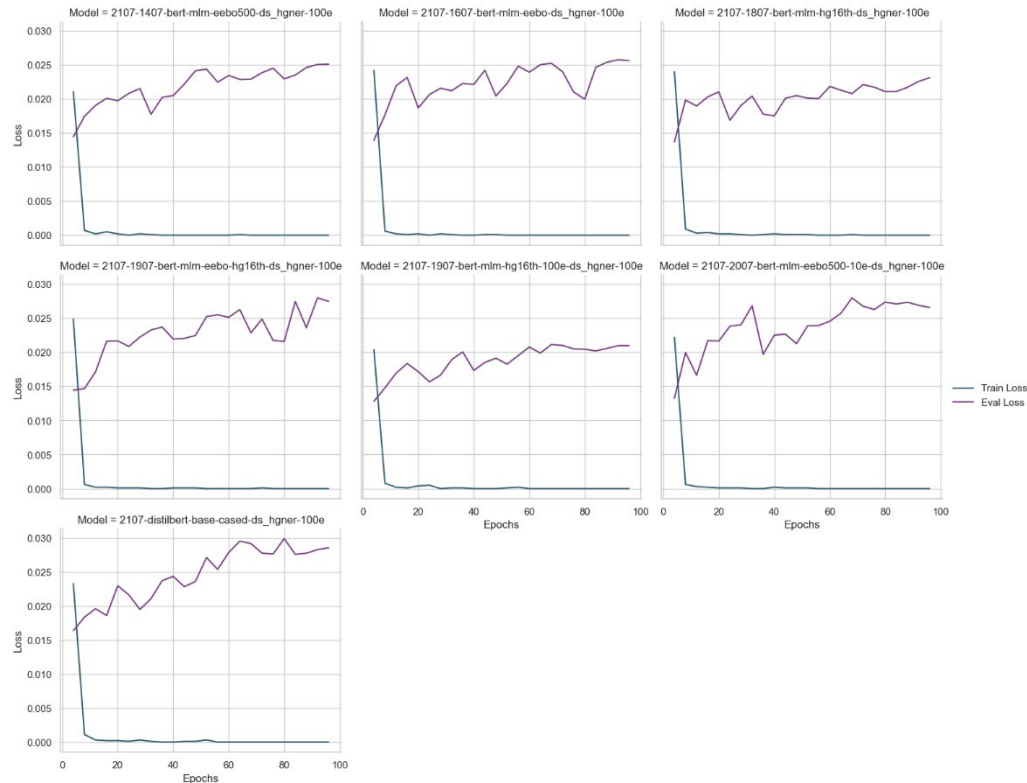
Universität
zu Köln



5. Methodology

- “For fine-tuning, most model hyperparameters are the same as in pre-training, with the exception of the batch size, learning rate, and number of training epochs.” (Devlin et al., 2019, p.2)
- Standard Practice for BERT (see Devlin et al., 2019)
 - **Epochs:** 2, 3, 4 **Batch Size:** 16, 32, 48 **Learning Rate:** 5e-5, 3e-5, 2e-5
- MLM probability: 0.15 (= 15% of the tokens are masked during training)
- Small changes in the hyperparameters can have significant effects on the resulting models
→ **Hyperparameter Optimization** + Evaluating various models against each other
- Training on a consumer *NVIDIA GeForce RTX 2060 Super* with 8 GB of VRAM
(e.g., about 30 hours for three epochs of EEBO)

5. Methodology



After about **three epochs**, the model gets increasingly worse at generalizing.

Transformers tend to overfit rather quickly.
→ *Early Stopping is very important*

6. Results

The Issue of Evaluation

In contrast to Corpus Linguistics, results (often models) in Computational Linguistics **must** be evaluated against some metric in order to assess their quality!

Masked Language Model Standards of Evaluation

- Perplexity (PPL) Test
 - Not applicable here as PPL is not well defined for BERT-like (MLM) models
- *GLUE* (Wang et al., 2018) and *SuperGLUE* (Wang et al., 2019) benchmarks
 - Sets of language tasks, offering a single-number metric to summarize model performance
 - The existing tasks/tests are based only on contemporary English!
- For historical models we would need to come up with a new set of evaluation tasks to use as a benchmark (i.e., a carefully collected corpus of (labelled) test data).

6. Results

Evaluation – Our Approach

Mask Filling Models

As there is no standard benchmark for historical data, our quantitative approach is reasserted by an additional qualitative approach:

- **Quantitative** Approach

Binary (Is the MASK in the predictions?)

Min. Distance $\sum_{i=0}^{no_examples} \min\{lev(mask, pred_1), \dots, lev(mask, pred_j)\}$

Weighted Distance $\sum_{i=0}^{no_examples} inv_rank * lev(pred_i, mask)$

Max. spaCy Sim. $\sum_{i=0}^{no_examples} \max\{sim(mask, pred_1), \dots, sim(mask, pred_j)\}$

- **Qualitative** Approach (two raters rank the models' performance on a suite of masked test sentences)

NER

- Precision = What proportion of predicted labels was actually correct? (Focus on FP)
- **Recall** = What proportion of actual labels was classified correctly? (**Focus on FN**)
- F1 score = weighted average of Precision and Recall
- Accuracy is not reported due to scarce amount of person tags and large amount of non-entity tags

$$p = \frac{TP}{TP + FP} \quad F1 = 2 * \frac{p * r}{p + r}$$

$$r = \frac{TP}{TP + FN}$$

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



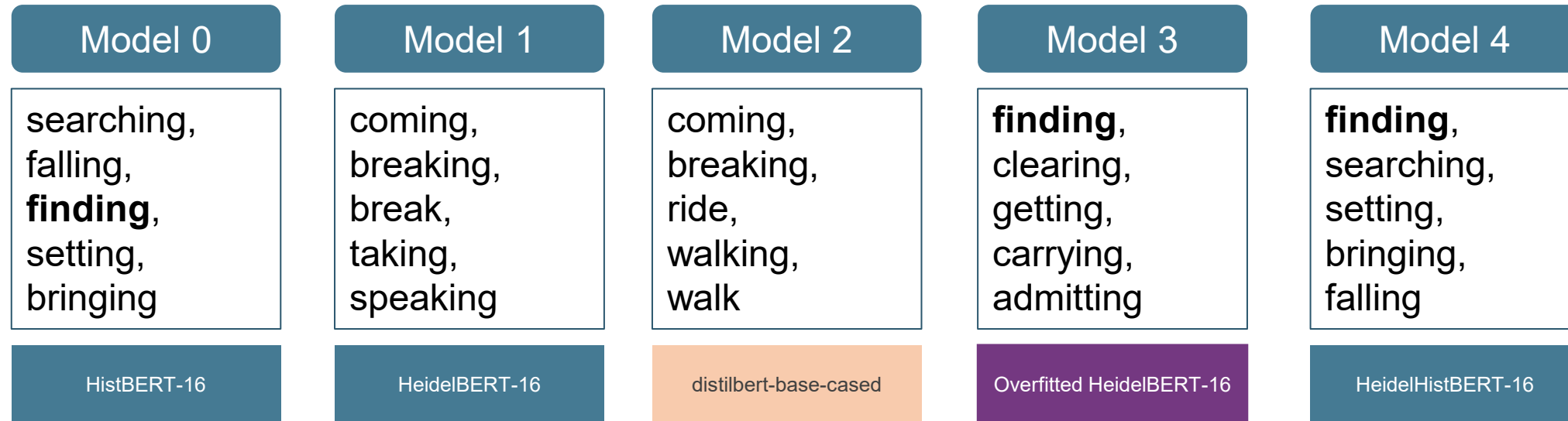
6. Results

Qualitative MLM Evaluation Process

➤ We find that language models create their own 'idiolects'!

“Plato in his platform for the [MASK] out of iustice hath two great vantages of me.” (Mulcaster, 1582, p. 232)

[MASK] = “finding”



Note: It's not the 'goal' of the MLM to find the exact match.

6. Results

MLM Scores

Performance across **101 examples** MLM tasks.

* For HeidelBERT-16 and HeidelHistBERT-16 there is some overlap between training/validation data.

Model	Binary	Weighted Distance	Min. Distance	Max. spaCy <i>en_core_web_lg</i> Sim.	Humans
HistBERT-16	36	396	297	54.37	Rank 3
HeidelBERT-16*	23	545	394	48.26	Rank 4
<i>HeidelBERT-16-UB</i>	24	658	381	48.31	NaN
HeidelHistBERT-16*	41	215	276	57.32	Rank 1
<i>HeidelHistBERT-16-UB</i>	40	157	292	57.1	NaN
Overfitted HeidelBERT-16*	48	184	247	62.33	Rank 2
distilbert-base-cased	18	375	421	44.56	Rank 5
distilbert-base-uncased	17	524	409	44.87	NaN

More is Better

Less is Better

Less is Better

More is Better

Higher (1) is Better

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



6. Results

Wherefor as in depenesse of meditation I drew like to **Plato**, tho in depth of iudgement but his fleting follower: So in order of deliuerie I depart from him and vtter my wares by retailing parcels, which he did ingrosse: when I had considered the generall ascending method of all learning, which while it is in getting, mounteth vp by degrees, but when it is gotten, doth sprede through out the state as sinews, veins, and arteries do through a naturall bodie, and withall maintains the state in full proportion of his best being, no lesse then the other do maintain the bodie, me thought I did perceiue some great blemish in the hole bodie of learning, as **Plato** no doubt, in the ripping vp, of right did find to be in fouernment. And as **Plato** himself by his own teaching did confirm his own precepts, whereby he brought forth a nuber of rare men, as euen the sharpe **Aristotle** & the eloquent **Demosthenes**, and by his singular plat of chosen gouernment, tho not all waie pleasing our religion and practis, did direct the bestconceits of the most studious people:

(Mulcaster, 1582, p. 233)

bert-base-NER

HG Reference

Both Models

I-PER and **B-PER**

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



6. Results

These foresayde sixe kyndes, **Quintilian** and other put vnder Prosopopoeia. Topographia, the discription of a place, as of Carthago in the fyrste of Eneidos. Hyther referre Cosmographie. Topothesia, the faynyng of a place, When a place is descrybed, as paraduenteure suche none is. Exaample of this is the Vtopia of **Syr Thomas Moore**. Or elles is not suche a place as it is, fayned to bee. As, is hell, and heauen in the syxte of Eneidos. Hyther pertayneth the situaciō of starres, in **Aratus**, **Higinus**, **Manilius** and **Pontanus**. Chronographia, the discription of tyme, as of nyght in the fowerth of Eneidos. Of the peace worlde in the fourth Egloge of **Virgil**. Of the foure ages in the fyrste of Metamorphoseos.

(Sherry, 1577, p. 47)

bert-base-NER

HG Reference

Both Models

I-PER and **B-PER**

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



6. Results

These foresayde sixe kyndes, **Quintilian** and other put vnder Prosopopoeia. Topographia, the discription of a place, as of Carthago in the fyrste of Eneidos. Hyther referre Cosmographie. Topothesia, the faynyng of a place, When a place is descrybed, as paraduenteur suche none is. Exaample of this is the Vtopia of **Syr Thomas Moore**. Or elles is not suche a place as it is, fayned to bee. As, is hell, and heauen in the syxte of Eneidos. Hyther pertayneth the situaciō of starres, in **Aratus**, **Higinus**, **Manilius** and **Pontanus**. Chronographia, the discription of tyme, as of nyght in the fowerth of Eneidos. Of the peace worlde in the fourth Egloge of **Virgil**. Of the foure ages in the fyrste of Metamorphoseos.

(Sherry, 1577, p. 47)

bert-base-NER

HG Reference

Both Models

All Possible Tags

Funded by

DFG

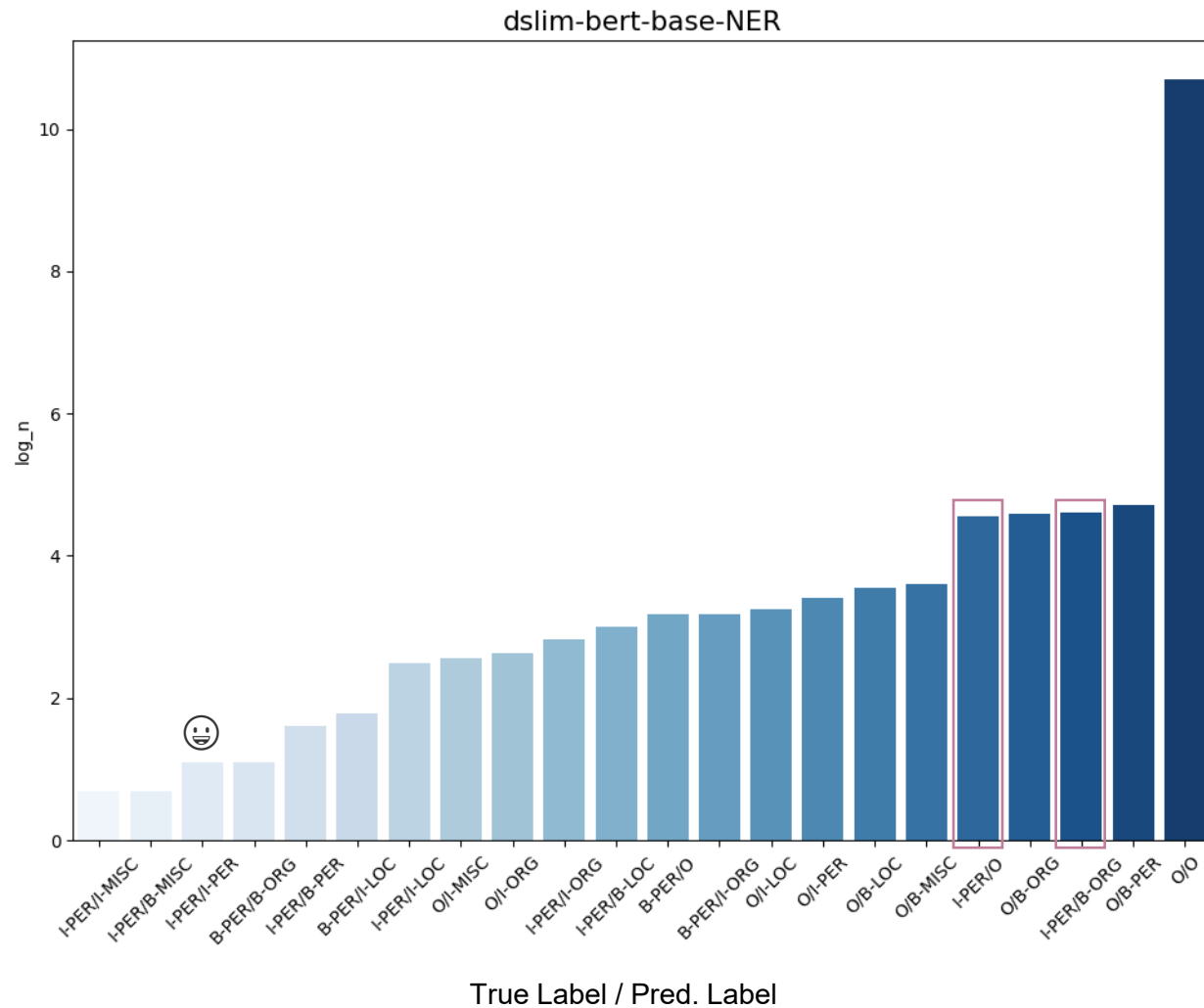
Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



6. Results

dslim/bert-base-NER



Modern English Reference

Metric	Value
F1	0.01313
Precision	0.02013
Recall	0.00974

Strict
(e.g., I-PER/B-PER)
need to match

Conflated into one tag (NER)

Metric	Value
F1	0.03463
Precision	0.0537
Recall	0.02556

Not Strict

Funded by



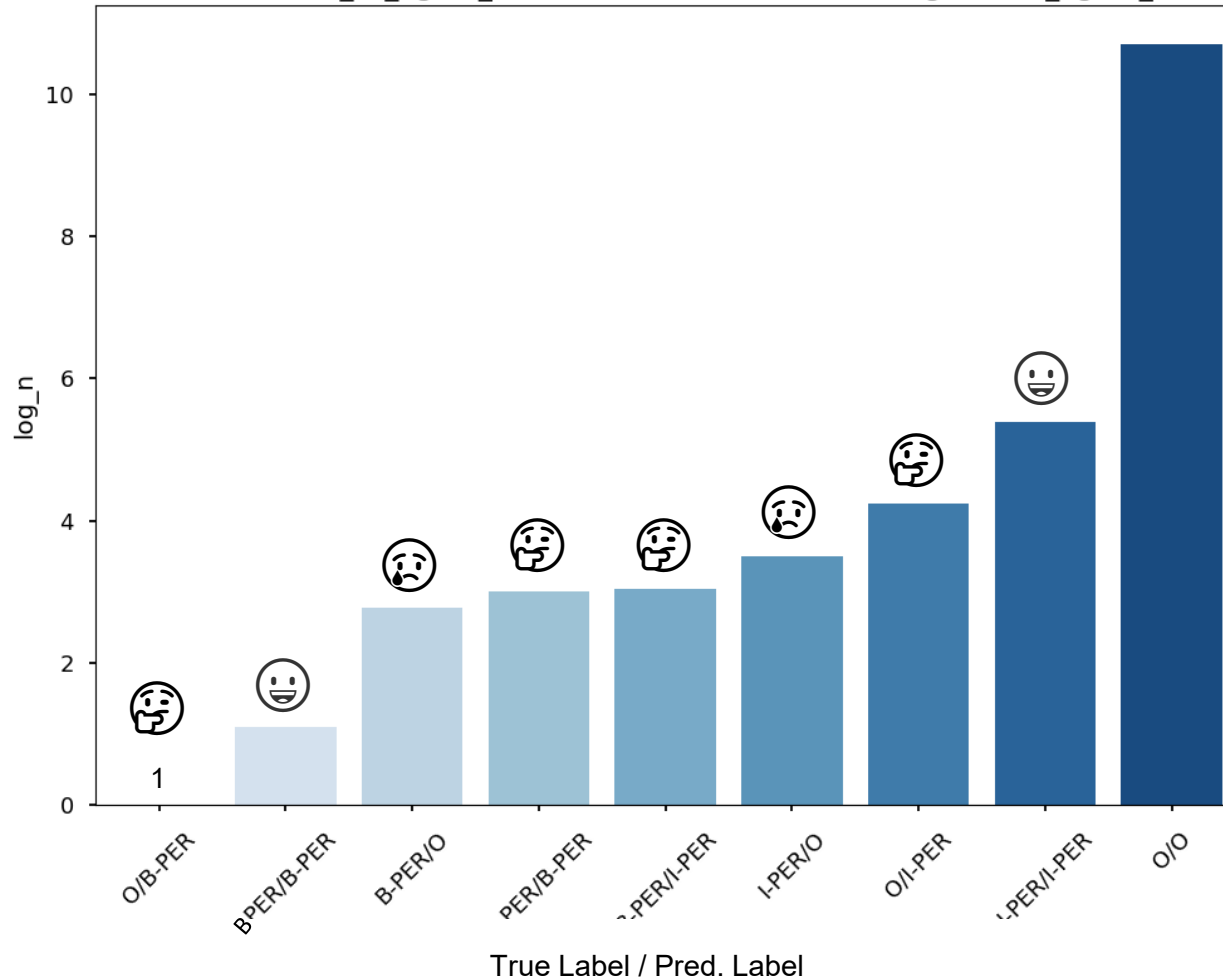
Deutsche
Forschungsgemeinschaft
German Research Foundation



6. Results

HeidelHistBERT-16

COLL-2207-train-ner_ds_hgner_8115-3e-1907-bert-mlm-eebo-hg16th-ds_hgner_8515-3e



Our 'best' model

Metric	Value
F1	0.77047
Precision	0.71856
Recall	0.83045

Strict
(e.g., I-PER/B-PER)
need to match

Conflated into one tag (NER)

Metric	Value
F1	0.81607
Precision	0.79042
Recall	0.84345

Not Strict

Funded by



Deutsche
Forschungsgemeinschaft
German Research Foundation



7. Learnings About the Methodology

Data / Corpus

- Meaningfully chunking (e.g., sentence segmentation) historical texts is hard but would be helpful for modelling.
- Tokenization, especially as a stepping stone before sub-word tokenization, has an impact on the models.
- Keeping the balance between leaving the data as is (training on the actual data) vs. pre-processing for better results
- There most likely is not enough historical data – even in EEBO – to reasonably train a transformer model from scratch (+ high cost money/environment) → Use a pre-trained BERT/ DistilBERT as our base

Training

- These models tend to overfit very quickly (monitoring + early stopping is important)
- The specific input requirements (e.g., sequence length, padding, special tokens) need to be taken into consideration
- The tokenizer needs to exactly fit the base model that's being fine-tuned (e.g., sub-word tokenization using *WordPiece* for DistilBERT)

7. Learnings About the Methodology

MLM

- Case seems to have less impact on model performance than we would have expected
- There are little established evaluation metrics for historical MLM models
→ *GLUE* and *SuperGLUE* are problematic due to the underlying tasks/datasets

Token Classification / NER

- If non-entities are classified as such, this needs to be taken into account when looking at evaluation metrics (e.g., accuracy)
- When applying modern models such as *DistilBERT* on historical it is beneficial to consider all entity tags as relevant
- Precise IOB tagging/classification (e.g., I-PER, B-PER) as in CoNLL2003 is hard to achieve; treat them as one

7. Conclusion and Next Steps

Conclusions

- Historical linguistics can benefit from historically informed language models.
- The performance of models on specific NLP tasks (e.g., Token Classification) can be increased by fine-tuning a modern LM on historical (domain) data.
- Nevertheless, models only trained on contemporary English can, in some cases, be successfully used on historical data.
- As NER is a specialized case of Token Classification it is reasonable to assume that these findings can be applied to, for example, Part-of-Speech tagging as well.

7. Conclusion and Next Steps

Next Steps

- Evaluating and improving this work especially using better/more sophisticated pre-processing as well as tokenization
- Applying these as well as other new models on the 17th, 18th, and 19th century components of HeidelGram
- **Future Downstream Tasks:** Better OCR-correction using MLMs, Text Classification, Tagging/Parsing, ...
- **Future Infrastructure Development:** Integrating ML/DL pipelines with our database
- **Open Science:**
 - After additional experimentation, we plan to release our fine-tuned models as well as our training code to the community
 - Discussing standardized evaluation metrics for historical models
- **Future of Language Modelling:**
 - This is a rapidly developing discipline and there are other projects on the horizon (e.g., Open GPT-X in Europe¹) which are developing powerful new models
 - Challenges: energy consumption, time, and financial constraints

Thank you for your
attention!



Bibliography I

Alexander Thamm. (2021, July 5). *Digitale Souveränität für Europa – Gemeinschaftsprojekt „Open GPT-X“ unter Mitwirkung der Alexander Thamm GmbH gewinnt Förderwettbewerb*. Retrieved August 2, 2021, from <https://www.alexanderthamm.com/de/blog/open-gpt-x-projekt-mit-alexander-thamm-gmbh/>.

Busse, B., Gather, K., & Kleiber, I. (2018). Assessing the connections between English grammarians of the nineteenth century: A corpus-based network analysis. In E. Fuß, M. Konopka, B. Trawiński, & U. H. Waßner (Eds.), *Grammar and corpora 2016* (pp. 435-442). Heidelberg University Publishing.

Busse, B., Gather, K., & Kleiber, I. (2019). Paradigm shifts in 19th-century British grammar writing: A network of texts and authors. In B. Bös & C. Claridge (Eds.), *Norms and conventions in the history of English* (pp. 49-71). John Benjamins.

Busse, B., Gather, K., & Kleiber, I. (2020). A corpus-based analysis of grammarians' references in 19th-century British grammars. In A. Cermakova & M. Malá (Eds.), *Diskursmuster - Discourse Patterns: Vol. 20. Variation in time and space: Observing the world through corpora* (pp. 133-172). Mouton De Gruyter.

Busse, B., Kleiber, I., Dumrukic, N., & Du Bois, S. (2021, July 13). *A corpus-based network analysis of 16th-century British grammar writing*. [Conference session]. Corpus Linguistics International Conference 2021, University of Limerick and Mary Immaculate College, Ireland.

Cameron, D. (2012). *Verbal hygiene*. Routledge.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Technologies*, 1:4171-4186.

Early English Books Online (EEBO) TCP. (n.d.). Retrieved November 24, 2020, from <https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/>.

Eisenstein, J. (2019). *Introduction to natural language processing*. The MIT Press.

Funded by
DFG Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



Bibliography II

Goyal, P., Pandey, S., & Jain, K. (2018). *Deep learning for natural language processing: Creating neural networks with python*. Apress.

Honnibal, M., Montani, I., van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo.

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 3rd. Edition. Stanford University.

Lhoest, Q., von Platen, P., Villanova del Moral, A., Wolf, T., Jernite, Y., Thakur, A., Tunstall, L., Patil, S., Drame, M., Chaumond, J., Plu, J., Davison, J., Brandeis, S., Le Scao, T., Sanh, V., Canwen Xu, K., Patry, N., McMillan-Major, A., Schmid, P., . . . Lagunas, F. (2021). *huggingface/datasets: 1.10.2*. Zenodo. Retrieved July 30, 2021, from <https://doi.org/10.5281/zenodo.5121423>

McCarthy, M. (2020). *Innovations and challenges in grammar: Innovations and challenges in applied linguistics*. Routledge.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). *A BERT-based transfer learning approach for hate speech detection in online social media*. arXiv preprint: <https://arxiv.org/pdf/1910.12574.pdf>.

Ruder, S. (2019). *Neural transfer learning for natural language processing* [Doctoral dissertation, National University of Ireland, Galway].

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint: arXiv:1910.01108.

Schweter, S., & Baiter, J. (2019). Towards robust named entity recognition for historic German. In I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, & M. Rei (Eds.), *Proceedings of the 4th workshop on representation learning for nlp (repl4nlp-2019)* (pp. 96–103). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4312>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv preprint: <http://arxiv.org/pdf/1706.03762v5>

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation

Universität
zu Köln



Bibliography III

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353-355, Brussels, Belgium, November. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 3266-3280.

Wei, T., Qi, J., & He, S. (2021). *A flexible multi-task model for BERT serving*. arXiv preprint: arXiv:2107.05377

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics*, Online, 38-45. Retrieved July 30, 2021, from <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [Review Article]. *IEEE Comput. Intell. Mag.*, 13(3), 55-75.

Zöllner, J., Sperfeld, K., Wick, C., & Labahn, R. (2021). Optimizing small BERTs trained for German NER. arXiv preprint: arXiv:2104.11559.

Bibliography IV

Corpus Data

Bullokar, W. (1586). *Brief grammar for English*. Edmund Bollifant.

Cootte, E. (1596). *The English schoole-maister teaching all his schollers, of what age soever, the most easie, short, and perfect order of distinct reading, and true writing our English-tongue, that hath euer yet beene knowne or published by any*. Printed by the Widow Orwin, for Ralph Jackson, and Robert Dextar

Meurier, G. (1586). *The coniugations in Englishe and Netherdutche, according as Gabriel Meurier hath ordayned the same, in Netherdutche, and Frenche*. Thomas Basson.

Mulcaster, R. (1582). *The first part of the elementarie which entreateth chefelie of the right writing of our English tung*. Thomas Vautroullier.

Sherry, R. (1577). *A treatise of the figures of grammer and rhetorike*. Ricardi Totteli.