

# Domain Adaptation with Linked Encyclopedic Data: A Case Study for Historical German

Thora Hagen

Institut für Deutsche Philologie, Julius-Maximilians-Universität Würzburg, Germany

## Abstract

This paper outlines a proposal for the use of knowledge graphs for historical German domain adaptation. From the *EncycNet* project, the encyclopedia-based knowledge graph from the early 20th century was borrowed to examine whether text-based domain adaptation using the source encyclopedia's text or graph-based adaptation produces a better domain-specific model. To evaluate the approach, a novel historical test dataset based on a second encyclopedia of the early 20th century was created. This dataset is categorized by knowledge type (factual, linguistic, lexical) with special attention paid to distinguishing simple and expert knowledge. The main finding is that, surprisingly, simple knowledge has the most potential for improvement, whereas expert knowledge lags behind. In this study, broad signals like simple definitions and word origin yielded the best results, while more specialized knowledge such as synonyms were not as effectively represented. A follow-up study was carried out in favor of simple contemporary lexical knowledge to control for historicity and text genre, where the results confirm that language models can still be enhanced by incorporating simple lexical knowledge using the proposed workflow.

## Keywords

language models, knowledge graphs, encyclopedic knowledge, semantics

## 1. Introduction


Based on Ryan's *principle of minimal departure* [30], our understanding of any text is highly dependent on our previous knowledge of the world. Consequently, depending on the type of text, for example a technical paper, we would also need to be experts in the same scientific field to be able to follow the arguments made. Another example would be historical literature, where certain cues in the text could only be understood by having a solid foundation on societies, fashion, or politics (among other topics) of that exact time period. The same can be argued for language models (LMs). When working with texts of a specific topic, type, genre, time period, etc., the language model's performance is also dependent on whether the training data matches the domain of the task at hand. In the case of digital humanities where, depending on the research domain, large text corpora may not be as readily available in comparison to contemporary English, the domain representation within the language model may not be stable enough. When employing a LM, researchers can either turn to a specialized pre-trained LM for the domain if available (e.g., MacBERT<sub>h</sub> [26] for historical English), or they have to perform domain adaptation of a general domain LM.

---

CHR 2024: Computational Humanities Research Conference, December 4 – 6, 2024, Aarhus, Denmark

✉ thora.hagen@uni-wuerzburg.de (T. Hagen)

ORCID 0000-0002-3731-6397 (T. Hagen)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This paper explores how an encyclopedia-based knowledge graph (KG) can be used to adapt language models specifically for historical German, with a focus on injecting the knowledge from that period. The goal is to demonstrate a simple workflow for researchers in the digital humanities to infuse LMs with domain knowledge using a KG. Especially in the humanities, there may be specialized resources available, for example dictionaries, thesauri, or lexicons, which can be transformed into knowledge graphs (see for example projects LiLa<sup>1</sup> and PURA<sup>2</sup>). KGs provide another form of knowledge representation aside from text, and they generally offer a wider variety of adaptation methods than text can. In this paper, the focus lies on the comparison of text and KG.

Specifically, this paper is concerned with the following research questions:

- How does adding a KG based on one encyclopedia as training data of a LM compare to simply adding that exact encyclopedia, i.e., is creating a KG worth it for creating a knowledge infused LM?
- What kind of knowledge shows the most improvement when injecting an encyclopedic KG into a LM (factual, lexical, linguistic)?
- Is a historical encyclopedia suited for historical domain adaptation?

For the experiment, two German encyclopedias from the early 20th century were chosen – one for training (*Meyers Großes Konversations-Lexikon* [27], dated in 1905, in the following referred to as *Meyers*) and one for evaluation (*Brockhaus Kleines Konversations-Lexikon* [4], dated in 1911, in the following referred to as *Brockhaus*). The former has been transformed into a semantic knowledge graph by *EncycNet*.<sup>3</sup> In a follow-up study, a comparison is also made between injecting contemporary linked semantic data, namely WordNet, and encyclopedic KGs in terms of improving lexical semantic relations in LMs.

## 2. Related Work

### 2.1. Knowledge Enhanced Pre-trained Language Models

The idea to inject language models with knowledge graphs belongs to the research area of knowledge enhancement. Generally speaking, not every form of knowledge can be learned by feeding vast amounts of continuous texts to a transformer model. Missing information, meaning explicit grounding in the real world [7], is not only apparent for domain knowledge (expert knowledge about e.g. drugs and diseases) but common sense and factual knowledge as well [3]. As an example, the newest development of OpenAI to include images and other media for model training (GPT-4o) seeks to tackle the grounding problem as well.

Knowledge enhanced pre-trained language models (KEPLMs) are language models that have been tuned to accommodate a specific area of knowledge better. While algorithmic adaptation is possible, many methods for creating KEPLMs rely on additional structured knowledge to inject the LMs with. These can be, among others, additional text snippets describing concepts

---

<sup>1</sup><https://lila-erc.eu/>

<sup>2</sup><https://pric.unive.it/projects/pura/home>

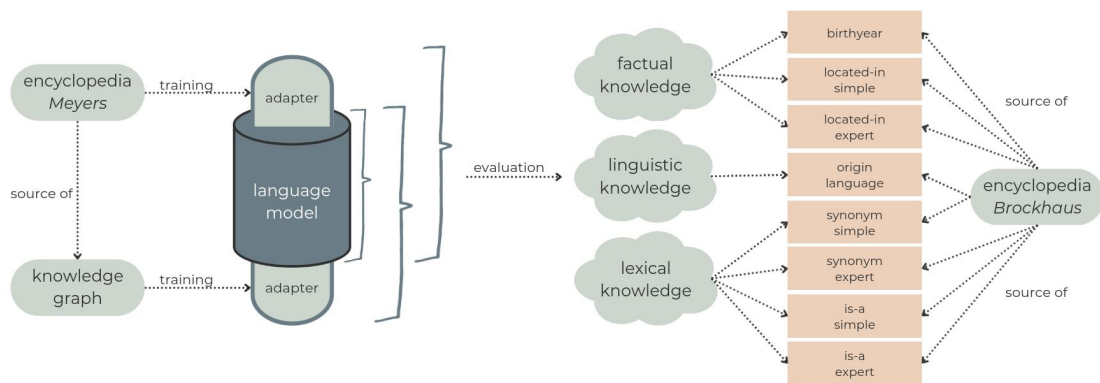
<sup>3</sup><https://encycnet.github.io/>; RDF knowledge graph available at <http://dx.doi.org/10.5281/zenodo.10219192>

or entities (e.g. dictionary definitions), tables, syntax trees, triples, rule systems, or knowledge graphs [15, 41]. Knowledge graphs bear an advantage over other structured data forms: They may be reshaped to other data structures and are thus highly flexible regarding the choice of method, and they can represent any type of human knowledge, meaning methods devised to accommodate knowledge graphs are flexible to adapt to any knowledge type.

Five different categories for knowledge enhancement using KGs can be broadly distinguished [28]. The first category is concerned with adapting the masked language modeling (MLM) training procedure (during pre-training or through continued training) using KG data. Firstly, the information given in the KG can be used to employ strategic masking during training (e.g., to mask multi-word expressions [35], assign masking probabilities for words through the graph structure [43], or mask head and tail entities when appearing in the same text passage [32], etc.). Secondly, the graph can be used to create new corpora through randomwalks [17], which can be used for MLM the same way natural continuous text can. The second category deals with employing additional tasks, either during pre-training or fine-tuning, which also use the KG as training data. These tasks can be, for example, creating stable knowledge graph embeddings [40], or predicting head, relation or tail of triples from the KG [29]. The third category attends to input fusion of KG and text, either by merging text into graph [34], graph into text [23], or merging features from the graph into the input layer of the transformer model [20]. These three categories have in common that they all aim to change the parameters of the language model. The final two categories of KEPLMs use KGs at inference (retrieval augmented generation) [24], or use the KG as evaluation data for interpretability and probing matters [36], where in both cases the language model keeps its original parameter configuration.

An additional trend for KEPLMs is the usage of adapters. First introduced by [39] as K(nowledge)-adapters, adapters are a set of layers introduced to the transformer model, where during training, only the parameters of the adapters are changed, while the rest of the LM stays frozen. This is meant to minimize "forgetting", where the original knowledge learned during pre-training gets overwritten, and thus ensures that the injected knowledge stays independent. In that way, multiple knowledge types can be injected into the model without interfering with each other or the original model (e.g., as per [39], factual and linguistic adapters).

In the following study, the focus lies on randomwalk generation as well as using adapters for training. Randomwalks have been previously employed for knowledge injection for a multitude of knowledge domains and tasks: factual and common sense knowledge [17], eventuality modeling [42], entity classification and link prediction for the biomedical domain [37], as well as lexical, medical and factual knowledge graph completion [21]. The method has also previously been employed to create taxonomic word embeddings [16]. The intuition of the approach lies in the assumption that traversing randomwalks in a graph can effectively capture its entire topology and map its contents into latent space (node2vec algorithm [10]). The randomwalk injection method was preferred here, as it allows for a fair comparison between the encyclopedia enhanced and knowledge graph enhanced language models. As the graph is deconstructed into text form, both can be created with continued MLM training, and only the input representation (continuous text vs. randomwalks) is different.



**Figure 1:** Schematic representation of the experimental design.

## 2.2. Domain Adaptation

The field of KEPLMs shares significant overlap with the research area of domain adaptation. As already briefly mentioned, domain adaptation is concerned with retroactively fitting general pre-trained LMs to a domain-dependent task. Some of the approaches in creating KEPLMs are quite similar, which is when structured knowledge is used to retroactively adapt a LM instead of influencing pre-training or inference. In domain adaptation, similar methods are for example continued MLM pre-training [11] or employing different masking strategies [2].

While these fields share the aspect of subsequent model fitting, KEPLMs prioritize using structured input data regardless of domain. Much work in this area focuses on improving factual or common sense related knowledge, not least because this is where most of the structured resources are digitally available (most importantly Wikidata and ConceptNet). Domain adaptation focuses more on solving the domain specific task, regardless of additional input used. This paper seeks make this connection explicit and set an example for the combination of the two fields, namely using a KG to adapt a LM to the historical German knowledge domain.

## 3. Infusing Language Models with Historical Encyclopedic Knowledge

### 3.1. Workflow Overview

A schematic representation of the proposed workflow can be found in Figure 1. As [17, 42, 37, 21] have demonstrated, randomwalks can be used to infuse LMs with new information, or, in the case of [16], even be the sole information source to build type embeddings. In this paper, the method for randomwalk creation was borrowed and adapted from [17]. All triples were extracted from *Meyers'* knowledge graph, where the predicates were resolved to simple German. As the original graph uses Wikidata properties, their German aliases were used for the verbalization (e.g. "P5973" to "Synonym"). [16] have shown that a non verbalization worked the best in their case, however as LMs process whole sentences, this simple verbalization method

was chosen here instead. The triples were parsed with networkX, and node2vec was used to create the walks. The procedure was slightly adapted from [17]: More unspecific relations, particularly "related to," were assigned a lower edge weight to reduce their probability of being selected during walks. Additionally, multiword expressions were not combined with underscores in this case. In total, 752,230 randomwalks were created. Examples can be found in Table 4 in the Appendix.

As a starting point, the current German state-of-the-art for encoder-decoder based models, gBERT-large,<sup>4</sup> was used, and an adapter using the LoRA [14] configuration was added. For this training setup, this means that only 0.234% of the original parameter size had to be trained (about 786K instead of about 335M). Using the encyclopedia's original text (see examples in Table 5), one adapter was trained on the MLM task. Then, another adapter was trained separately on the randomwalk KG representation of the same encyclopedia. Both adapters were trained using the same hyperparameters each, which are 8 epochs, MLM probability of 0.15, and learning rate of 1e-4. Additionally, the model's perplexity [31] during the randomwalk training was calculated on a sample of the OSCAR dataset (used for the pre-training of gBERT) over the course of 24 epochs (see Appendix 4). Here, it can be seen that even though the use of an adapter should mitigate forgetting pre-trained knowledge, the perplexity increases quite steadily for OSCAR. However, it also declines for the randomwalks, confirming that the model is improving on this dataset during training. This shows that there is still a trade-off, and training with the randomwalk corpus should not be extended beyond a certain point, which is why the training was stopped at epoch 8.

The evaluation procedure relies on predicting the correct word from a given word plus word relation using the fill-mask pipeline. The creation of these word pair datasets is described in the following section. Using the [MASK]-token and a verbalization of the expected relation, the LM is prompted to predict the second word of the pair. Some examples can be found in Table 6 in the Appendix. Then, the performance is calculated by the correct hits within the top predictions of the LM. Other evaluation methods focus on embedding extraction of word types by fusing the token embeddings from multiple sentences and measuring the relationship via cosine distance. As the embedding method could be volatile to sentence sampling, and could potentially conflate the different dimensions of word "closeness" through just cosine distance, the evaluation strategy used here seeks to negate the randomness through sampling and takes the nuances of word relations into account.

### 3.2. Creation of the Evaluation Dataset

The evaluation consists of probing the original and knowledge infused LMs on different knowledge types, which is meant to assess which information can actually be ingested with the proposed workflow: factual, linguistic, and lexical semantic knowledge. Several different datasets consisting of word pairs were constructed to cover these three tasks. The word pairs were extracted with regular expressions from another German encyclopedia of the same time period (*Brockhaus*) to make sure that the historical variation and text genre of the input encyclopedia also matches the evaluation data. The decision to use two different encyclopedias for training and testing stems from the nature of the task, which is not about generalizing knowledge,

---

<sup>4</sup><https://huggingface.co/deepset/gbert-large>

but rather about learning specific, encyclopedic relations such as synonyms and factual associations. Unlike more general language tasks, the relations captured in an encyclopedia – especially those pertaining to domain-specific knowledge – are inherently difficult to generalize beyond their specific context. By testing on a second encyclopedia, *Brockhaus*, the aim is to evaluate how well the model has internalized and can retrieve the learned relationships rather than generalizing abstract patterns; a similar approach to earlier model "semantic retrofitting" methods [8, 33].

For the evaluation data, 5 different types of word pair lists were constructed: people and their year of birth,<sup>5</sup> places and where they are located, words and their language of origin, pairs of synonyms, and definitions of concepts (also referred to as *is-a* relation or hypernyms). The first two datasets represent factual knowledge, the third dataset represents linguistic knowledge, and the last two represent lexical semantic knowledge.

However, the content of encyclopedias in general is not only historical, but at times extremely detailed, as they do not only cover general knowledge but a lot of domain specific knowledge as well, such as chemistry or botany, for instance. Similarly, some facts are easier than others depending on how well known the entity in question is. Where possible, the datasets were separated into two splits: simple and expert knowledge. For the dataset about places, the population size along with the location were extracted. All places with a (historical) population size exceeding 70,000 were added to the simple knowledge category. Places with a population size between 30,000 and 70,000 were counted as expert knowledge. For both lexical semantic datasets, GermaNet [12] was used to gauge the level of specificity of the word pairs. In more precise terms, the corresponding synset was retrieved for the second word of the pair and along with it its level in the hierarchy of GermaNet terms. From a psychological point of view, a higher hypernym depth in the hierarchy would correspond with a higher specificity / expert knowledge, while a more shallow depth would insinuate a simpler kind of knowledge. When given more than one synset for one word, the minimum depth of these synsets was chosen. The extracted hypernym depths of synonyms exhibit a mean of 7.71 (median of 7), while the *is-a* pairs have a mean of 6.52 (median of 6). The difference is to be expected, as the definitions should always indicate an upper hierarchical level in contrast to the synonyms. When comparing the encyclopedia synonyms to another dataset commonly used for evaluating word-level similarity (German translation of SimLex [19]), the "expertness" of the encyclopedia becomes apparent. The mean depth of SimLex word pairs is 5, meaning that on average, SimLex pairs are 2 hierarchy levels above encyclopedia pairs (see distribution comparisons in Figure 2).

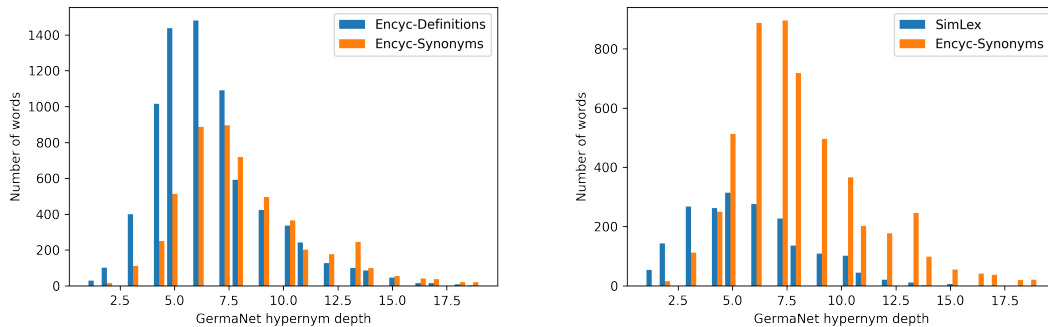
As a result, all word pairs where the predicted word has a hypernym depth of 6 or lower were categorized as simple, and 7 and up counted as expert.<sup>6</sup> For the linguistic and year of birth datasets, the data were not split because no immediate additional feature for separation could be identified. All created datasets can be found on github.<sup>7</sup>

---

<sup>5</sup>The birth dates in the *Brockhaus* dataset exhibit a median of 1811 and are highly skewed, with a long tail extending back to the year 1000. The 25th percentile (Q1) is 1757, and the 75th percentile (Q3) is 1835.

<sup>6</sup>While two separate thresholds could have been introduced here, a single "expertness" threshold ensures a reliable comparison across both datasets. It is based on the assumption that a lexeme's "expertness" level should not change depending on whether it appears in a synonym or hypernym context.

<sup>7</sup><https://github.com/ThoraHagen/HistED/>



**Figure 2:** Side-by-side hypernym depth distributions of three selected datasets. Left: Comparison between two *Brockhaus* datasets – synonyms and definitions. Right: Comparison between two synonym datasets – SimLex and *Brockhaus* synonyms.

**Table 1**

Performance comparison of original gBERT-large model and models enhanced with encyclopedic knowledge across all datasets. *ency* refers to the encyclopedia full-text adapted model. *ency-KG* refers to the encyclopedic knowledge graph adapted model.

task	y.o.b.	loc simple	loc expert	lang	is-a simple	is-a expert	syn simple	syn expert
# instances	8943	226	487	11956	4384	2961	890	1817
method	avg. dist.	hits@10	hits@10	hits@3	hits@10	hits@10	hits@10	hits@10
gBERT	163	0.14	0.06	0.43	0.06	0.04	0.06	0.02
ency	151	0.15	0.06	0.56	0.08	0.05	0.07	<b>0.03</b>
ency-KG	<b>100</b>	<b>0.18</b>	<b>0.07</b>	<b>0.66</b>	<b>0.13</b>	<b>0.08</b>	<b>0.08</b>	<b>0.03</b>

### 3.3. Results

The results of both encyclopedia and encyclopedia-KG adapted models can be found in Table 1. For evaluation, hits@n were calculated across all datasets except for *year of birth*, where the average prediction error (distance from the true year) was used instead. The hits@n metric reflects the proportion of correct answers ranked in the top  $n$  model fill-mask predictions, where higher values indicate better performance. Overall, the KG-based domain adaptation is able to outperform the full-text adaptation, though the degree of improvement varies across tasks.

**Factual Knowledge** Both evaluation datasets representing factual knowledge exhibit some improvement, albeit minor. As indicated above, for the *year of birth* dataset, the evaluation focused on the average difference between the actual year of birth and the top 3 predicted years, because the hits@n metric for all three models yielded near-zero scores, even when  $n$  was set to large values. This approach provides a better sense of how close the models’ predictions were to the correct year, given the low performance in ranking accuracy. It can be seen that even though the model makes a somewhat better educated guess (as in ”in an

encyclopedia published in 1905 there should not be any birthdates mentioned before that”) as the average distance is reduced by about 50 years, the precision is still poor. A qualitative review of the prompts did not find any correlation between correct guesses and a person’s fame (as fame may reflect both simple and expert knowledge in this case). Further work on quantifying fame and splitting the dataset accordingly is necessary to confirm this notion. A similar sentiment can be observed with the *location* datasets. Both simple and expert location knowledge exhibit minor improvements of about 1-4 percentage points (pp.) more hits@10. One possible explanation could be that the majority of information about locations is already contained through the pre-training of gBERT-large, and not many evaluation examples contain information that changed until today. The location dataset is quite fine-grained, meaning that rather than countries, smaller regions are given as the true label, which also affects the exact prediction accuracy. A qualitative examination of some evaluation instances show that more sensible location predictions were made overall, even if the exact label is not predicted (see Appendix 6). However, similar to the *birthyear* dataset, the accuracy is very low.

**Linguistic Knowledge** The task for assigning the origin language to a word represents linguistic knowledge in this setup (for example *Absolut* and Latin). Because the outcome space of the prediction is presumably much more limited than for the other datasets, the evaluation setup was narrowed to hits@3. The observed improvements for both gBERT ency and gBERT ency-KG are quite high, with 13 and 23 pp. more hits respectively. Out of all datasets, the improvements are the highest here. However, it needs to be addressed that this dataset is quite imbalanced, as most true labels are either French, Latin, or Greek, meaning that the improvements seen could just be the nature of a language distribution shift. Other words with a different language of origin may not be predicted as well. In terms of a historical domain adaptation, it still can be said that the method performs as intended: It is more likely that a word in a German historical encyclopedia stems from one of these three languages, which is exactly what the dataset reflects.

**Lexical Semantic Knowledge** While for both synonym and *is-a* relations some improvements can be observed, the two datasets perform quite differently. Firstly, the *is-a* relations outperform the synonyms, with about 7 pp. more hits for the former and merely 2 pp. more hits for the latter concerning the simple relations. Secondly, both lexical expert variants fall behind their simple counterparts, with only about a 4 pp. difference for definitions and a 1 pp. difference for synonyms. Both results indicate that simpler lexical knowledge is more beneficial to language models than expert lexical knowledge. One could assume that the simpler knowledge would already be contained in gBERT through the OSCAR dataset pre-training, and that the injection would benefit the representation of specialized knowledge more, so this is a surprising result.

In summary, it can be said that 1) factual knowledge shows a trend towards improvement but lacks the specificity that these two datasets demand, 2) linguistic knowledge shows greater improvements, however this result may stem from a simple distribution shift, 3) lexical knowledge shows greater improvements for the upper hierarchy level of *is-a* relations while synonyms are



harder to predict. Across the datasets but especially for lexical knowledge, simple knowledge still bears more room for improvement, while expert knowledge is harder to ingest. This may seem surprising, as previous studies have demonstrated that language models already possess, or have largely mastered, basic semantic knowledge [25, 6].

## 4. Lexical Semantic Knowledge for LM Infusion

In this section of the paper, the focus therefore lies on confirming whether LMs still have room for improvement for contemporary lexical semantic knowledge by removing two confounding factors from the previous experiment: original text type (encyclopedia) and historicity. This is why instead of the encyclopedia KG, WordNet is used for LM injection in the following experiment.

Concerning KEPLMs, many studies have been conducted that evaluated mostly on factual knowledge, task-based common sense or domain knowledge, and lexical-based studies are rather rare (see [15] for an overview of recent KEPLM studies). To the best of available knowledge, there are no studies that explicitly evaluate the upper limit of lexical semantic knowledge improvement for randomwalk-fitted LMs. Similar studies focusing on lexically informed LMs are most importantly LIBERT [18] and Mirror-BERT [22]. LIBERT introduces a new classification loss during pre-training based on whether a given tuple holds a semantic relation using WordNet plus Roget’s Thesaurus. The authors evaluate on the GLUE benchmark, where the focus lies on sentence level semantics, as well as the lexical simplification task, a variant of assessing word level similarity using context from sentences. Mirror-BERT does not rely on external data but instead introduces text corruption, where the model learns to cluster true and false (corrupted) text samples. Evaluation is based on sentence level and word level tasks, including word level similarity.

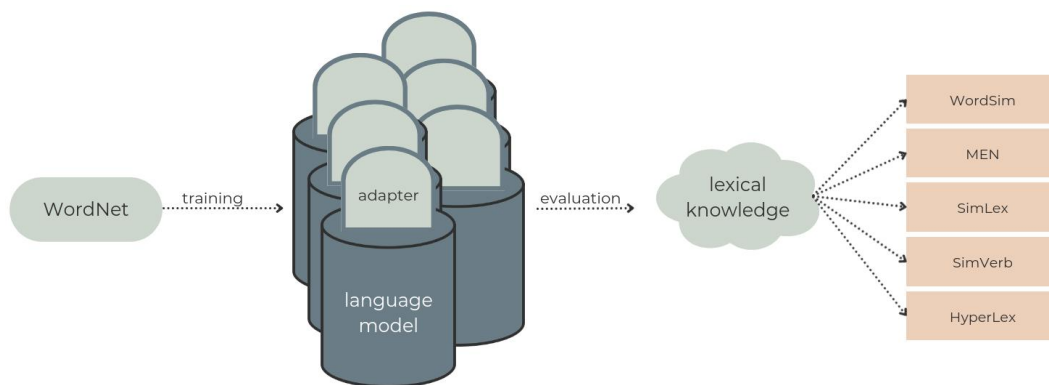
Different from LIBERT and Mirror-BERT (aside from the injection method), this section also takes different model sizes into account and evaluates on three different lexical tasks: association, similarity and entailment.

### 4.1. Methodology

A visualization of the WordNet workflow can be found in Figure 3. First, all triples were extracted from the WordNet database. In a second step, the relations were verbalized to mimic natural language, e.g. ”synonym” to ”is a synonym of.” Again, the verbalized triples were parsed with networkX and the node2vec algorithm was used to create 258,239 randomwalks. Some examples of WordNet randomwalks can again be seen in Table 4 in the Appendix.

To evaluate the retrofitting effectiveness for lexical semantics in particular, five datasets were chosen as stand-ins for three different lexical semantic tasks: SimLex [13] and SimVerb [9] for evaluating semantic word similarity, WordSim [1] and MEN [5] for semantic word relatedness, and HyperLex [38] for evaluating lexical entailment. All datasets are score-annotated word pairs, e.g. on a scale of 0 to 10, *happy* and *cheerful* score a similarity of 9.55 (SimLex).

Because these datasets represent the strength of one semantic relationship between two words rather than a binary relation, the evaluation method was slightly adapted in this experiment. Similar to before, by using the fill-mask strategy, the evaluation focuses on probing each



**Figure 3:** Workflow of the WordNet experiment.

**Table 2**

Overview of language models in the WordNet study. Rows: Text size used for pre-training. Columns: Number of parameters.

	66-88M	110M-125M	340M-355M	8B
<b>50GB</b>	DistilBERT-base	BERT-base	BERT-large	
<b>160GB</b>	DistilRoBERTa-base	RoBERTa-base	RoBERTa-large	
<b>83TB</b>				Llama-3-8B-instruct

language model on relation prediction given the first word of each pair. However, the top 100 words are predicted here as compared to the previous experiment. Here, the inverse indices of all word pair matches are compared to the true dataset scores using Spearman’s correlation. For example, the RoBERTa-large model predicts *cheerful* from the task “happy is a synonym of <mask>” at rank 5, which would translate to a similarity score of 95. In other words, scores are assigned to word pairs by their prediction ranking of each model.

Multiple models of different parameter sizes as well as pre-training text-sizes are compared to assess how the method scales with these model differences. Similar to the experiment before, only LoRA-adapters were trained instead of the whole model. In comparison to the BERT-family of encoder-decoders, Llama-3<sup>8</sup> was also included as a point of reference for large language decoder-only models. To match the evaluation strategy of predicting a single word, the instruct variant (specialized on adapting to user generated tasks) was chosen over the chat variant (specialized on text generation). Here, an adapter was prompt-trained to predict the object given a subject and predicate statement using WordNet triples, similar to the fill-mask task of encoder-decoders (for a similar approach see [36]). An overview of all models can be found in Table 2. All WordNet adapters were trained three separate times to mitigate possible model unstableness<sup>9</sup> due to the random weight initialization, and the mean Spearman’s correlations

<sup>8</sup><https://huggingface.co/meta-Llama/Meta-Llama-3-8B-Instruct>. The model weights were cast to bfloat16 for memory efficiency (low precision training).

<sup>9</sup>The results exhibit a mean standard deviation of 0.01. Standard deviations were calculated per dataset and model.

**Table 3**

Spearman’s correlations of both original and WordNet-fitted models with the true scores of each selected dataset. Averages are reported for all WN models. Performance difference is denoted as  $\Delta$ . Bold: Highest overall performance for each task. Underlined: Highest  $\Delta$  for each task.

	WordSim	MEN	SimLex	SimVerb	HyperLex
DistilBERT-base	0.307	0.397	0.205	0.019	0.203
DistilBERT-base WN	<b>0.525</b>	<b>0.584</b>	0.434	0.217	0.366
$\Delta$	0.218	0.187	0.229	0.198	0.163
BERT-base	0.376	0.410	0.161	0.048	0.231
BERT-base WN	0.505	0.566	0.434	0.216	0.474
$\Delta$	0.129	0.156	0.273	0.168	0.243
BERT-large	0.256	0.333	0.151	0.013	0.249
BERT-large WN	0.522	0.551	0.530	0.316	0.530
$\Delta$	<u>0.266</u>	<u>0.218</u>	<u>0.379</u>	<u>0.303</u>	0.281
DistilRoBERTa-base	0.216	0.248	0.106	0.003	0.105
DistilRoBERTa-base WN	0.407	0.443	0.221	0.146	0.288
$\Delta$	0.191	0.195	0.115	0.143	0.183
RoBERTa-base	0.271	0.331	0.093	0.005	0.195
RoBERTa-base WN	0.442	0.482	0.440	0.299	0.478
$\Delta$	0.171	0.151	0.347	0.294	0.283
RoBERTa-large	0.386	0.423	0.409	0.219	0.274
RoBERTa-large WN	0.504	0.506	<b>0.629</b>	<b>0.481</b>	<b>0.590</b>
$\Delta$	0.118	0.083	0.224	0.262	<u>0.316</u>
Llama-3-8B-instruct	0.223	0.242	0.100	0.005	0.049
Llama-3-8B-instruct WN	0.281	0.285	0.236	0.157	0.310
$\Delta$	0.058	0.043	0.136	0.152	0.261

across these three adapters per model are reported.

## 4.2. Results

The results of the WordNet adapted models can be found in Table 3. For all the models, the injection of WordNet is able to benefit the representation of lexical semantics using a randomwalk-adapter.

Overall, a higher parameter size is beneficial for this approach, not only in terms of the generally best performing models, but highest performance jumps as well. The two word relatedness tasks, WordSim and MEN, do better on models trained with less text, while word similarity and lexical entailment do better on the models trained with more text. An indicator for this kind of separation could be the clearness of the evaluated relation: While semantic relatedness indicates the degree of association between two words, semantic similarity indicates the degree of synonymy and lexical entailment the degree of hypernymy. Compared to the latter two, the former is a much more fuzzy concept. This could indicate that with increasing parameter size, more refined relations can be better represented instead of just word association. In the case

of hypernymy, which is not a symmetrical relation compared to the other two tasks, RoBERTa-large has the largest overall performance and largest performance difference. The same trend can be found in the non-fitted versions of the models. For RoBERTa-large, both similarity and entailment are already represented significantly better in comparison to the mean of the other models, while the performance on relatedness is comparable to the others. Concerning the Llama model, even though it is also showing signs of improvement, it cannot compare to the encoder-decoder-based models in this setup. A similar trend like for the large models shows however, which is that on average, associations show the least improvement, followed by similarity, and finally entailment benefits the most. The contrast to the other models may stem from the differences in model pre-training and not necessarily because of size differences only. Further studies will be needed to explore how to better tailor the lexical adapter approach to decoder-only models.

The results indicate that more refined tasks, here lexical entailment, benefit more from the increased model size, while the less precise association task shows more stagnation across different model sizes. In terms of the pre-training corpus size, the results are less intuitive. The distilled variant of RoBERTa does not show any significant advantage over its BERT counterpart. For the base variant, again, only synonyms and entailment show minor improvements over BERT-base. When using WordNet randomwalks for creating a lexically informed LM, it can be seen that models with more parameters benefit from the method for synonym and entailment relations. Corpus size may only matter when both parameter and text size are comparatively high. For word association, the performance differences are generally not as high and the task shows a negative correlation with original corpus size. The assumption that larger models already contain the majority of lexical knowledge and do not benefit from lexical injections is therefore not true, and the results align with previous studies in this regard [18, 22].

## 5. Summary and Outlook

In summary, this paper has shown that extracting a KG from a resource can be helpful when domain-adapting a general LM to a historically informed LM. The models were evaluated with a two-dimensional approach: one categorical dimension for the type of knowledge (factual, linguistic, lexical semantics) and another binary dimension to distinguish simple from expert knowledge. The main finding is that surprisingly, simple knowledge still bears the most potential for improvement while expert knowledge falls behind. The WordNet follow-up study confirmed that language models can still be enhanced with simple lexical knowledge.

Regarding the question of whether a historical encyclopedia is suited for historical domain adaptation, it can be said that that it depends on the use case of the language model. Encyclopedias contain specialized knowledge because the diverse fields of expertise discussed, as well as the historical perspective, directly influence its lexical richness. Thus, encyclopedias contain specialized knowledge also in terms of expert semantic relations. When using the approach discussed here, one should target more precisely what kind of knowledge to inject. Using the entire knowledge graph may not send strong enough signals for fitting a specific task. Here, broad signals such as simple definitions and language of word origin showed the

best results, while synonyms especially could not be represented as effectively. Employing a knowledge graph, future work could therefore explore multiple ways of limiting the training data to either specific relations (e.g. to target synonyms only) or historical knowledge domains. When controlling for these two confounding factors (expert domains discussed plus historical expertise) using WordNet, it can be observed that the same method is capable of injecting contemporary lexical knowledge such as synonymy into LMs, where even the larger models generally perform better.

In future work, concerning model analysis, the test suite for the encyclopedic evaluation will be diversified more. Currently, a binary classification of simple and expert knowledge, determined through an automatic approach using GermaNet, is being used. However, the dataset might exhibit a more intuitive notion of expert knowledge when manually annotating and deriving a continuous score from the annotations. Additionally, more relations will be added to the dataset to ensure that the results do not stem from peculiarities of the chosen relation and better represent the overall task.

There are more nuances to model training in this study that have not been taken into account yet. For one, the hyperparameters have been kept stable for the entirety of the experiments to ensure comparability between models. Potentially, this means that the upper bound of the KG injection models have not been reached. Another question to pursue would be how this method transfers to other tasks based on sentences. Instead of MLM adapters, the training of task-based adapters such as NLI is also possible. In future work, the evaluation could then also focus on how stacking both the KG adapter and another task-trained adapter (with both adapters activated during inference) could influence task performance. The hypothesis could be that certain tasks that rely on lexical information, such as sentiment prediction or semantic textual similarity, could also benefit from WordNet, for example. Finally, future work will also aim to better understand the differences between encoder-decoder and decoder-only language models. The disparities in pre-training (MLM vs. causal language modeling) may have significant impacts on infusing these models with more knowledge. Therefore, different injection strategies or prompting strategies will need to be compared to better assess the possibilities of knowledge-enhanced pre-trained LLMs.

## References

- [1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. “A study on similarity and relatedness using distributional and wordnet-based approaches”. In: *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*. 2009, pp. 19–27.
- [2] M. Aragon, A. P. L. Monroy, L. Gonzalez, D. E. Losada, and M. Montes. “DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 15305–15318.
- [3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity”. In: *Proceedings of the 13th International Joint Conference*

- on *Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 675–718.
- [4] F. A. Brockhaus, ed. *Brockhaus' Kleines Konversations-Lexikon*. 5th ed. Leipzig: Brockhaus, 1911.
- [5] E. Bruni, N. K. Tran, and M. Baroni. “Multimodal Distributional Semantics”. In: *Journal of Artificial Intelligence Research* 49 (2014), pp. 1–47.
- [6] T. A. Chang and B. K. Bergen. “Language model behavior: A comprehensive survey”. In: *Computational Linguistics* 50.1 (2024), pp. 293–350.
- [7] D. Coelho Mollo and R. Millièrè. “The vector grounding problem”. In: *arXiv preprint arXiv:2304.01481* (2023).
- [8] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith. “Retrofitting word vectors to semantic lexicons”. In: *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference i* (2015), pp. 1606–1615. DOI: 10.3115/v1/n15-1184. arXiv: 1411.4166.
- [9] D. Gerz, I. Vulić, F. Hill, R. Reichart, and A. Korhonen. “SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 2173–2182.
- [10] A. Grover and J. Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.
- [11] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 8342–8360.
- [12] B. Hamp and H. Feldweg. “GermaNet-a lexical-semantic net for German”. In: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 1997, pp. 9–15.
- [13] F. Hill, R. Reichart, and A. Korhonen. “SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation”. In: *Computational Linguistics* 41.4 (2015), pp. 665–695. DOI: 10.1162/COLI\_a\_00237.
- [14] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2021.
- [15] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li. “A survey of knowledge enhanced pre-trained language models”. In: *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [16] F. Klubička, A. Maldonado, A. Mahalunkar, and J. Kelleher. “English wordnet random walk pseudo-corpora”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 4893–4902.

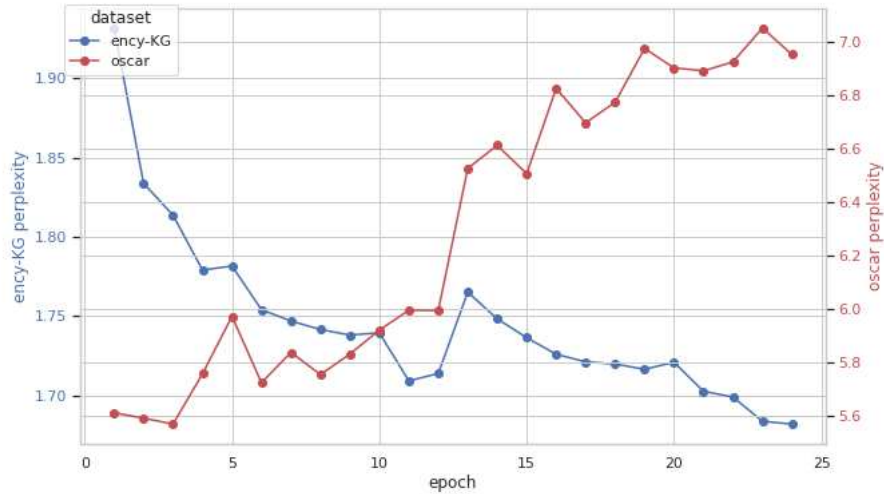
- [17] A. Lauscher, O. Majewska, L. F. Ribeiro, I. Gurevych, N. Rozanov, and G. Glavaš. “Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers”. In: *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 2020, pp. 43–49.
- [18] A. Lauscher, I. Vulić, E. M. Ponti, A. Korhonen, and G. Glavaš. “Specializing unsupervised pretraining models for word-level semantic similarity”. In: *arXiv preprint arXiv:1909.02339* (2019).
- [19] I. Leviant and R. Reichart. *Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling*. 2015. arXiv: 1508.00106 [cs.CL].
- [20] Y. Levine, B. Lenz, O. Dagan, O. Ram, D. Padnos, O. Sharir, S. Shalev-Shwartz, A. Shashua, and Y. Shoham. “SenseBERT: Driving Some Sense into BERT”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Online: Association for Computational Linguistics, 2020, pp. 4656–4667. doi: 10.18653/v1/2020.acl-main.423.
- [21] Q. Lin, R. Mao, J. Liu, F. Xu, and E. Cambria. “Fusing topology contexts and logical rules in language models for knowledge graph completion”. In: *Information Fusion* 90 (2023), pp. 253–264.
- [22] F. Liu, I. Vulić, A. Korhonen, and N. Collier. “Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 1442–1459.
- [23] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. “K-BERT: Enabling language representation with knowledge graph”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 2901–2908.
- [24] R. Logan, N. F. Liu, M. E. Peters, M. Gardner, and S. Singh. “Barack’s Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5962–5971. doi: 10.18653/v1/P19-1598.
- [25] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. “Dissociating language and thought in large language models”. In: *Trends in Cognitive Sciences* (2024).
- [26] E. Manjavacas and L. Fonteyn. “Adapting vs. pre-training language models for historical languages”. In: *Journal of Data Mining & Digital Humanities* (2022).
- [27] J. Meyer, ed. *Meyers Großes Konversations-Lexikon*. 6th ed. Leipzig: Bibliographisches Institut, 1905–1909.
- [28] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. “Unifying large language models and knowledge graphs: A roadmap”. In: *IEEE Transactions on Knowledge and Data Engineering* (2024).

- [29] Y. Qin, Y. Lin, R. Takanobu, Z. Liu, P. Li, H. Ji, M. Huang, M. Sun, and J. Zhou. “ER-ICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, 2021, pp. 3350–3363. DOI: 10.18653/v1/2021.acl-long.260.
- [30] M.-L. Ryan. “Fiction, non-factuals, and the principle of minimal departure”. In: *Poetics* 9.4 (1980), pp. 403–422.
- [31] S. Serrano, Z. Brumbaugh, and N. A. Smith. “Language Models: A Guide for the Perplexed”. In: *arXiv preprint arXiv:2311.17301* (2023).
- [32] T. Shen, Y. Mao, P. He, G. Long, A. Trischler, and W. Chen. “Exploiting Structured Knowledge in Text via Graph-Guided Representation Learning”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, 2020, pp. 8980–8994. DOI: 10.18653/v1/2020.emnlp-main.722.
- [33] R. Speer and J. Lowry-Duda. “ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 2017, pp. 85–89.
- [34] T. Sun, Y. Shao, X. Qiu, Q. Guo, Y. Hu, X.-J. Huang, and Z. Zhang. “CoLAKE: Contextualized Language and Knowledge Embedding”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 3660–3670.
- [35] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu. “Ernie: Enhanced representation through knowledge integration”. In: *arXiv preprint arXiv:1904.09223* (2019).
- [36] M. Sung, J. Lee, S. Y. Sean, M. Jeon, S. Kim, and J. Kang. “Can Language Models be Biomedical Knowledge Bases?” In: *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL). 2021, pp. 4723–4734.
- [37] Y. Tan, Z. Zhou, H. Lv, W. Liu, and C. Yang. “Walklm: A uniform language model fine-tuning framework for attributed graph embedding”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [38] I. Vulić, D. Gerz, D. Kiela, F. Hill, and A. Korhonen. “HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment”. In: *Computational Linguistics* 43.4 (2017), pp. 781–835. DOI: 10.1162/COLI\_a\_00301.
- [39] R. Wang, D. Tang, N. Duan, Z. Wei, X.-J. Huang, J. Ji, G. Cao, D. Jiang, and M. Zhou. “K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pp. 1405–1418.
- [40] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang. “KEPLER: A unified model for knowledge embedding and pre-trained language representation”. In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 176–194.



- [41] J. Yang, G. Xiao, Y. Shen, W. Jiang, X. Hu, Y. Zhang, and J. Peng. “A survey of knowledge enhanced pre-trained models”. In: *Journal of the Association for Computational Machinery* 37.4 (2023).
- [42] C. Yu, H. Zhang, Y. Song, and W. Ng. “CoCoLM: Complex Commonsense Enhanced Language Model with Discourse Relations”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. 2022, pp. 1175–1187.
- [43] T. Zhang, C. Wang, N. Hu, M. Qiu, C. Tang, X. He, and J. Huang. “DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 11703–11711.

## A. Perplexity



**Figure 4:** Perplexity progression during ency-KG training of gBERT-large on a sample of the OSCAR dataset versus the ency-KG dataset over the course of 24 epochs.

## B. Randomwalk Examples

**Table 4**

Two examples of randomwalks from WordNet and ency-KG (derived from *Meyers Großes Konversations-Lexikon*).

graph	randomwalk
WordNet	unmentionable is similar to impermissible. impermissible is similar to tabu. tabu is a synonym of proscribed. proscribed is a synonym of forbidden. forbidden is similar to impermissible. impermissible is similar to proscribed. proscribed is a synonym of prohibited.
WordNet	albuterol is a bronchodilator. bronchodilator is a medication. medication is a synonym of medicinal_drug. medicinal_drug is a synonym of medication. medication is a synonym of medicament. medicament is a synonym of medicinal_drug. medicinal_drug is a drug.
ency-KG	Kirrung verwandter Begriff Ankörnen. Ankörnen verwandter Begriff Blasenfüßer. Blasenfüßer Hyperonym gelbbraune Dracänenblasenfuß. gelbbraune Dracänenblasenfuß Hyponym Thrips. Thrips Definition Insektengruppe.
ency-KG	Synonymenwörterbuch verwandter Begriff Wörterbuch. Wörterbuch verwandter Begriff Handwörterbuch. Handwörterbuch verwandter Begriff Frerichs. Frerichs Synonym Friedrich Theodor Frerichs. Friedrich Theodor Frerichs geboren 24. März 1819.

## C. Encyclopedia Articles

**Table 5**

Two examples of articles from *Meyers Großes Konversations-Lexikon* used for training the full-text-based *ency* model. Only the first few tokens of the articles are displayed; see the full-text examples at "Blasenfüßer" and "Wörterbuch".

---

Blasenfüßer (Physopoda, Thysanoptera), Insektengruppe von sehr zweifelhafter Stellung im System, wird zu den Falschnetzflüglern gestellt und umfaßt winzige Tierchen mit zylindrischem Kopf, saugenden Mundwerkzeugen, sehr schmalen, stark befransten Flügeln, die bisweilen auch fehlen, und runden Hastscheiden statt der Klauen an den Füßen. Die B. leben auf Blättern, nehmen die zarte Oberhaut derselben weg und erzeugen dadurch oft bedeutenden Schaden. [...]

---

Wörterbuch (Lexikon), ein in rein alphabetischer oder alphabetisch-etymologischer Ordnung verfaßtes Verzeichnis von Wörtern und Eigennamen (welch letztere aber bisweilen fehlen oder ein besonderes W. bilden) mit oder ohne beigefügte Erklärung in der nämlichen oder in einer andern Sprache. [...]

---

## D. Example Predictions from the Fill-Mask Pipeline

**Table 6**

Examples of predicted word tokens (English translation only) from gBERT and gBERT encyc-KG for different tasks in the evaluation procedure (*Brockhaus Kleines Konversations-Lexikon*). The bracketed word is an example of the task, while the rest of the verbalization is kept the same for every test instance.

verbalization	model	top 5 predicted tokens	true label
(Leonian contract) is a [MASK].	gBERT	joke, other, good, compromise, proposal	social contract
	gBERT encyc-KG	contract, agreement, document, federal state, war	
(Maidstone) is located in [MASK].	gBERT	Wales, Scotland, South Africa, Great Britain, England	Kent
	gBERT encyc-KG	England, Scotland, Wales, Massachusetts, Great Britain	
(Toll) is a synonym of [MASK].	gBERT	toll, unjust, highway, infrastructure, fraud	customs
	gBERT encyc-KG	toll, fee, tax, fees, value-added tax	
(William George Armstrong) was born in year [MASK].	gBERT	1887, 1882, 1892, 1891, 1874	1810
	gBERT encyc-KG	1835, 1837, 1832, 1831, 1847	
(Accurate) is a word from the language [MASK].	gBERT	German, Japan, Italy, France, Latin	Latin
	gBERT encyc-KG	Latin, German, French, English, Italian	