Lexical Semantic Change Annotation with Large Language Models

Thora Hagen
Leibniz Institute for the German Language (IDS), Mannheim, Germany
University of Würzburg, Germany
hagen@ids-mannheim.de

Abstract

This paper explores the application of state-of-the-art large language models (LLMs) to the task of lexical semantic change annotation (LSCA) using the historical German DURel dataset. We evaluate five LLMs, and investigate whether retrieval-augmented generation (RAG) with historical encyclopedic knowledge enhances results. Our findings show that the Llama3.3 model achieves comparable performance to GPT-4o despite significant parameter differences, while RAG marginally improves predictions for smaller models but hampers performance for larger ones. Further analysis suggests that our additional context benefits nouns more than verbs and adjectives, demonstrating the nuances of integrating external knowledge for semantic tasks.

1 Introduction

The increasing application of natural language processing (NLP) methods to the humanities presents a range of challenges, particularly when working with historical or non-standard language data. One such challenge is the detection of lexical semantic change (LSC) (Tahmasebi et al., 2021; Periti and Montanelli, 2024), i.e. how words have shifted in meaning due to cultural, social, or linguistic contexts. Although large language models (LLMs) have demonstrated considerable success in modern language tasks, their ability to effectively interpret historical texts is still limited. Many works show how language models struggle with such long-tail knowledge (Kandpal et al., 2023; Wang et al., 2023). Linguistic limitations in particular then influence other computational research fields, such as the analysis of literary texts, where subtle shifts in meaning must be captured to correctly interpret a historical text via the lens of a historical reader.

LLMs are predominantly trained on contemporary data, which may lack the necessary historical linguistic context to accurately process and interpret older texts. Many tasks in the humanities that employ LLMs are directly or indirectly dependent on historical knowledge, including the recognition of historical facts, events, and discourse, as well as changes in word meaning over time (e.g., *gay*, *awful*, *computer*). When working with older text collections, this impacts applications such as historical sentiment analysis, the classification of historical texts, the analysis of narrative and character descriptions, and even machine translation of older documents. Understanding how different LLMs represent historical semantics is crucial for researchers who work with historical texts, as it informs their choice of model for specific tasks.

This paper investigates the performance of multiple state-of-the-art LLMs on the task of LSC annotation for historical German. The goal is to evaluate the models' ability to detect semantic shifts, and therefore to potentially infer which of the models would be the best to represent historical semantics via this proxy task. The research questions of this paper are the following: RQ1: How well can the current state-of-the-art LLMs annotate lexical semantic change given two contexts and a target word? RQ2: Can historical, referential knowledge increase the performance of the lexical semantic change annotation task?

2 Lexical Semantic Change Detection

Lexical semantic change detection (LSCD) is a well-established subfield of computational linguistics and NLP. Given a large diachronic text corpus (i.e. a corpus is that is divided into two or more time slices), the goal is to automatically detect which words have changed in meaning. These results are then compared to manually annotated gold datasets of word meaning shift, where annotations are either on a binary or graded (1-4) scale of relatedness (Tahmasebi et al., 2021; Kurtyigit et al., 2021).

Typically, word embeddings are used to detect

Proceedings of LaTeCH-CLfL 2025, pages 172–178

Table 1: One example of a context pair with target word highlighted in bold letters (engl. *the press* and *to press*) from the DURel dataset.

context 1	context 2	rating
V. Die Geschichte des Rechts der **Presse**. 1) Die Elemente der Geschichte.	[...] **Pressen** Sie mir kein offenherziger Bekenntniß ab. Jch liebe Sie, und bin ganz die Jhrige.	1 (Unrelated)

LSC; both contextualized approaches relying on transformer-based models or approaches based on static word vectors are possible. With static approaches, word embedding models of each corpus slice are created individually and are then aligned (e.g. through orthogonal Procrustes) (Hamilton et al., 2016; Wevers and Koolen, 2020). Contextualized approaches first calculate sets of token embeddings for every word in question, taking the surrounding context into account. Here, only one model is used, and one time slice corresponds to one set of token embeddings instead. The cosine distance of one word to itself in different time slices is then used to gauge its semantic "stableness", indicating whether the word has changed meaning or not. Contextualized approaches could either average embeddings beforehand or calculate average pairwise distances instead (Giulianelli et al., 2020; Laicher et al., 2021).

To evaluate LSCD approaches, some datasets were already manually created for different languages. Here, multiple contexts are strategically paired so that they contain the same word. Human annotators evaluate these pairs, assessing how stable the word meaning within these two contexts appears to be. Repeated annotation of these context pairs then results in a word usage graph, from which individual word senses or a single category of meaning shift can be inferred (Schlechtweg et al., 2018, 2020; Kurtyigit et al., 2021). In the remainder of the paper, this process will be referred to as lexical semantic change annotation (LSCA).

With the emergence of large decoder-only language models, such as ChatGPT, the view on LSCD has changed. Previously, embeddings were the only reliable way of detecting LSC, simply because not enough training data are available to fine-tune a model on the task. So, LSCD is currently a strictly unsupervised task. Now, LSCD can rely on the vast amount of knowledge that has been pre-trained into LLMs, which already show huge popularity with zero-shot (unsupervised) approaches. So far however, the results of using LLMs for LSCD have been mixed. Periti et al. (2024) compared BERT with GPT-3.5 for LSCD by having the models rank 37 target words by degree of change, finding comparable performance between the two. In contrast, Wang and Choi (2023) reported better performance when prompting GPT-4 to rate context pairs, outperforming the BERT-based embedding approach. However, their study was limited to the short-term change dataset, TempoWiC.

In this paper, the focus lies on LSCA, where we assess meaning change at the instance level by prompting LLMs to annotate context pairs, and we evaluate through correct classification, not ranking. By applying this approach to the long-range change, German dataset DURel, we aim to extend previous findings on the overall representation of semantic change in LLMs.

3 Experiments

3.1 Dataset

For this experiment we chose DURel (Schlechtweg et al., 2018), which is a manually annotated dataset for German LSCD. 22 target words were selected on the basis of previous intuitions that these words can represent change in meaning. Five annotators rated the context pairs on a scale of 1 to 4 (see example in Table 1). The contexts were derived from the DTA corpus (*Deutsches Textarchiv*), spanning roughly the 19th century.

3.2 Methods

First, we compiled all individual judgments of the context pairs in DURel. We averaged the scores for context pairs across the annotators and rounded the results, resulting in 1318 averaged use pairs (439 'identical', 413 'closely related', 303 'distantly related', 163 'unrelated'). Ratings of 0 ('unsure') were excluded. To assess the stability and generalizability of the experiment, rather than evaluating the model on the entire dataset once, we randomly sampled 30 instances 20 times, which allowed us to compute 20 F1 scores for model evaluation. The samples were randomized initially and then kept fixed across all experiments. This approach helps

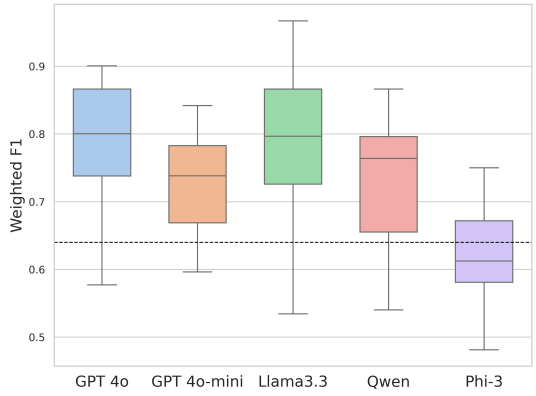

Figure 1: Binary LSCA results (F1) for 5 select models. Dotted line indicates majority baseline.

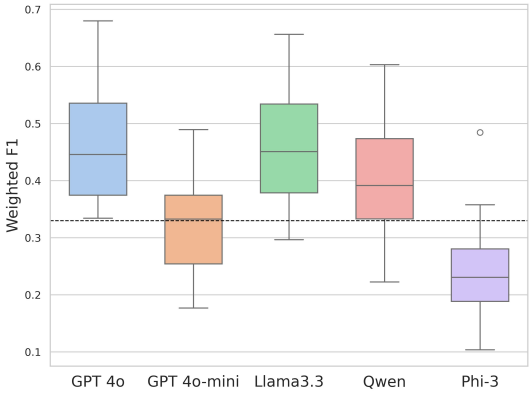

Figure 2: Graded LSCA results (F1) for 5 select models. Dotted line indicates majority baseline.

assess the consistency and reliability of the model's performance across different data subsets, offering a better indication of its overall effectiveness on new, unseen data.

Based on the results of Periti et al. (2024), the prompt was designed as a zero-shot approach, asking to rate the target word based on the two contexts for similarity (see Appendix A). DURel is constructed around comparing lexical items, not tokens, which means that target words may appear as different derivative or orthographical forms in two contexts. The prompt includes instructions not to base decisions on whether the same lexical item also happens to appear as the same token or not. Finally, the models are asked to first give a detailed explanation of their reasoning and then state their rating as one of the 4 relatedness categories, not their numerical equivalents. During the evaluation, the ratings were then extracted from the models responses with regular expressions. 5 different models were evaluated: GPT-4o-mini, GPT-4o, Llama3.3-70B, Phi3, and Qwen2.5-72B. With this selection of models, we mostly wanted to compare open vs. closed domain as well as larger vs. smaller LLMs.[1]

3.3 Results

For the results, we chose to look at the original 4-way classification as well as binary classification, where we labeled classes 1 and 2 as 'change' and 3 and 4 as 'no change', to simulate the two most prevalent evaluation strategies (Schlechtweg et al., 2020). As expected perhaps, the smaller models with 3B and 8B parameters (Phi and GPT-4o-mini) under perform compared to the larger models – for the case of Phi-3 even below the majority baseline (Fig. 1, 2). However, the Llama model demonstrates comparable performance to GPT-4o, despite the latter having 200B parameters. This trend is consistent across both binary and graded LSCA evaluations. The main differences are that the binary results exhibit higher volatility, which is mainly reflected in a larger first quartile, and the performance margin between large and small models is more noticeable for the graded task.

The relatively high spread of F1 scores across different sample sets suggests that model performance is highly dependent on the specific instances chosen. This variability implies that either certain target words or contexts are more challenging for the model or that the model struggles with consistent predictions. This highlights the need for more data annotations so that model performance can be evaluated on a more diverse and representative set of target words, reducing the impact of instance-specific variability and improving overall reliability.

3.4 Historical Prompt Augmentation

In this section, we evaluate whether providing additional lexical context may improve the LSCA task. Retrieval Augmented Generation (RAG) has widely been adopted because of its efficiency instead of fine-tuning when it comes to providing additional input to LLMs (Gao et al., 2023). We therefore turn to a historical German encyclopedia

[1]Embedding-based approaches are commonly used for LSCD, where the goal is to track semantic shift over time by comparing word distributions in different corpora. However, LSCA requires evaluating semantic change at the level of individual word instances in context. Since embedding models like BERT are not explicitly trained for this task and lack sufficient training data for reliable instance-level change detection, they are not directly applicable for LSCA.

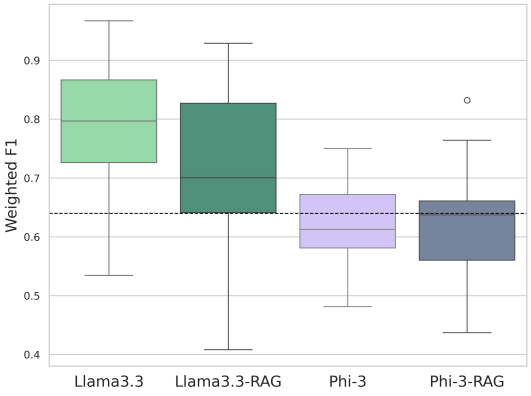

Figure 3: Binary LSCA results (F1) for Llama3.3 and Phi3 with their RAG equivalents. Dotted line indicates majority baseline.

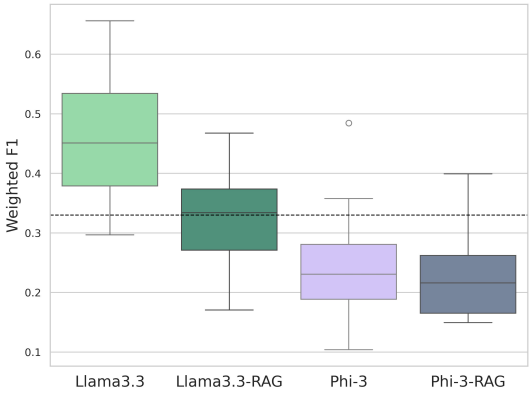

Figure 4: Graded LSCA results (F1) for Llama3.3 and Phi3 with their RAG equivalents. Dotted line indicates majority baseline.

of the early 20th century (Meyers' Großes Conversations Lexikon, 1905). We chose this resource for mainly two reasons: 1) encyclopedias generally contain a vast amount of referential knowledge that could help with providing more context to how a word was used and 2) this encyclopedia aligns with the DURel dataset time-wise. While LLMs are predominantly trained on contemporary data, the inclusion of an older encyclopedia provides the model with a historical perspective. This may allow the model to simultaneously process knowledge from the early 20th century and its own contemporary knowledge.

Instead of storing the entries in the RAG database as a whole, we have split the texts further in order to be able to carry out a more precise retrieval. We chose two approaches: context-sentence construction and triple construction. For the context sentences, we first split the encyclopedia corpus by sentence delimiters and added the entries' headword to every sentence as contextual information. One retrieved sentence is for example: *[Motivieren] in der Kunst, vor allem in der Dichtkunst: eine dargestellte Handlung oder Begebenheit [...].* Triples were extracted through diverse heuristics and regular expressions to condense the content of the encyclopedias further, e.g. *Motiv Synonym Beweggrund*. All texts are embedded using OpenAI's `text-embedding-3-small`. During inference, 5 instances per database per context (= 20 text instances) are retrieved as additional prompt information as follows:

Let the embeddings of a context sentence T and a target word W be represented as $\mathbf{e}_T$ and $\mathbf{e}_W$, respectively. The calculation of the final embedding $\mathbf{e}_{\text{final}}$ is given by:

$$\mathbf{e}_{\text{final}} = \frac{\mathbf{e}_T + w \cdot \mathbf{e}_W}{\|\mathbf{e}_T + w \cdot \mathbf{e}_W\|}$$

where: w is a fixed weight ($w = 1.5$ in our case), and $\| \cdot \|$ denotes the Euclidean norm. This means that for retrieval, more weight is given to the target word in relation to the surrounding context, so that similarity mostly considers the target word and is not as much influenced by other words in the context. The nearest neighbors are calculated from either database via cosine similarity.

Take, for instance, the target word *Vorwort*, which could mean both "preface"[2] or "preposition."[3] The retrieval produced *Preface Definition Vorrede* and *[Vorwort] Auch Vorrede eines Buches (praefatio)* as the two most relevant texts for the former, as well as *Vorwort verweist auf Präposition* and *[Präposition] (lat.), Vorwort, ein Redeteil, der entweder dem von ihm regierten Worte vorausgeht, z. B. mit Vernunft, oder, was seltener ist, nachfolgt, z. B. des Vaters wegen.* for the latter. In this case, even though the target word is the same, the query correctly produces differently contextualized documents for the two meanings, demonstrating that the approach is viable. This additional context is then integrated into the prompt as well (see Appendix A). The information is described as optional, meaning that the model should also assess whether the information is helpful or not.

[2]Context from DURel: "[...] und sprach im Vorworte ganz wie ein guter Landsmann der beiden Dänen. [...]"

[3]Context from DURel: "[...] und die Verhältnisse durch Vorwörter ausdrückt. So z. E. kann man anstatt hier, heute, rechts, bald, rc. sagen"

3.5 Results

The results of the RAG approach are mixed: For the smaller Phi3 model, some improvements could be observed while the approach for the larger model actively impairs model performance (Fig. 3, 4). This could be due to the fact that these models already capture the same kind of historical knowledge and additional context only provides noise. To better understand these changes, we analyzed the transition from the previous models to the RAG models at the level of individual target words, examining whether the new predictions correctly or incorrectly leaned towards similarity or dissimilarity.

Generally speaking, we find that the changes meant higher similarity predictions after RAG, and re-classification affected certain words disproportionally. Consequently, most errors occurred due to a higher similarity prediction, where especially the words *feine*, *flott*, and *packen* (*fine, fast, to pack*) were affected. These words accounted for 22 out of 49 new mistakes in this category. Now correctly assigned similarity scores due to higher similarity are mostly the words *Kinderstube*, *Anstellung*, und *Bilanz* (*nursery, employment, balance*). Changes towards dissimilarity mostly and erroneously affected *locker* (*loose; loosely*), while correct reclassification in the dissimilarity category seems evenly spread (but less likely overall).

It could be hypothesized that ingesting the encyclopedia generally benefits the contextualization of nouns rather than verbs and adjectives, also given the fact that encyclopedias typically focus on explaining concepts, while verbs and adjectives may not receive the same level of detailed explanation.

Furthermore, we observe that language models tend to overemphasize domain differences when annotating LSC, which remains largely unchanged even with the addition of RAG. For instance, in the case of *englisch*, the model classifies *englische Krankheit* ("English disease") and *Englische Flotte* ("English fleet") as having distantly related meanings explicitly due to being used in different domains, despite both usages fundamentally referring to England. While the two contexts indeed belong to different domains—medicine versus military—the core meaning of *englisch* remains stable. This suggests that the model relies heavily on contextual domain differences rather than recognizing the underlying semantic continuity of a word. The fact that this pattern persists with RAG indicates that additional historical context does not necessarily correct this bias, highlighting a potential limitation in how LLMs process lexical meaning across different domains.

4 Summary of Findings

We can conclude that the open-domain Llama3.3 model performs on par with the closed-domain GPT-4o model for the LSCA task (though larger models tend to perform better overall), suggesting that both models contain similar knowledge of historical semantics. Providing additional context through a historical encyclopedia yields mixed results: the augmentation only slightly positively impacts the smaller model, and the performance is highly dependent on the target word. Overall, we found that LLMs may process semantics differently than humans would, as the models put a larger emphasis on the domain in which a word is used. Future work will need to address this challenge, uncovering the reasoning for the models behavior as well as steering models more towards a human intuition of lexical semantics.

5 Limitations

This paper presents only a preliminary experiment on how to further explore the LSCA task using LLMs. First, the prompts could be further refined, potentially incorporating more of the original annotation guidelines from the DURel dataset. Similarly, the handling of additional context through RAG could be optimized, including adjustments to how the retrieved information is presented to the model. The preprocessing of encyclopedias for the RAG database, as well as the choice of retrieval strategy, could also be improved. In this study, several parameters were kept stable for pragmatic reasons—such as the weighting of the target word for embedding creation, the number of items retrieved, and the decision to retrieve two separate contexts—but these should be further evaluated and tuned in future experiments. Finally, the results and observed volatility highlight the need to study a larger set of target words and to explore how different external sources might influence the automatic annotation of lexical semantic change.

References

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen

Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany.

Joseph Meyer, editor. 1905. *Meyers Großes Konversations-Lexikon*. Leipzig.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6985–6998.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In *Proceedings of the 16th Conference of the European Chapter of the ACL: Student Research Workshop*.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (Chat) GPT v BERT Dawn of Justice for Semantic Change Detection. In *Findings of the ACL: EACL 2024*, pages 420–436.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. 2021. *Computational approaches to semantic change*. Language Science Press, Berlin.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Ruiyu Wang and Matthew Choi. 2023. Large language models on lexical semantic change detection: An evaluation. *arXiv preprint arXiv:2312.06002*.

Melvin Wevers and Marijn Koolen. 2020. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4):226–243.

A Prompt Template

This is the prompt template used for both experiments. The parts highlighted in blue were used for the RAG experiment only:

"You are a knowledgeable linguistic assistant with expertise in lexical semantics. Your task is to analyze the meanings of words in different contexts to determine how related they are. You will be given a target word and two sentences/contexts where this word appears. Note that the word may not appear in the same token form in both contexts; it could be a different lexical form of the same word (e.g., an inflected or derived form). Carefully assess both the similarities and differences in meaning without assuming that they must be different. Additionally, you will be provided with information: 1. Knowledge graph triples related to each context. 2. Relevant encyclopedia sentences where the target word appears. These sentences may provide extra insights or cultural background but should not be used as the primary basis for comparison.

1. Carefully analyze each context independently to determine the target word's meaning. 2. Compare the meanings directly, focusing on their **core semantic similarities or differences**. 3. If the meanings are **exactly the same** or **completely different**, prioritize "Identical meanings" or "Unrelated meanings" over intermediate ratings. 4. Use the knowledge graph triples and encyclopedia sentences as **clarification tools**, but do not let overlaps influence your judgment unfairly.

Rating: - Identical meanings: The word's meaning is exactly the same in both contexts. - Closely related meanings: The word's meanings are very similar, with only minor differences in nuance or usage. - Distantly related meanings: The word's meanings are somewhat connected but show clear differences in usage or interpretation. - Unrelated meanings: The word's meanings have no apparent connection between the contexts.

Input:
Target Word: '{target_word}'
Context 1: '{sentence_1}'
Context 2: '{sentence_2}'
Relevant Knowledge Graph Triples for Context 1: {triples_1}
Relevant Encyclopedia Sentences for Context 1: {encyclopedia_sentences_1}
Relevant Knowledge Graph Triples for Context 2: {triples_2}

Relevant Encyclopedia Sentences for Context 2: {encyclopedia_sentences_2}

Output: Explanation: [Provide a detailed explanation of the relatedness of the target word in both contexts. Use the encyclopedic data as supplementary insights, but focus on comparing the meanings of the word as used in Context 1 and Context 2.] Rating: [Identical meanings, Closely related meanings, Distantly related meanings, or Unrelated meanings]"