Synonymy Representations Compared and Evaluated Across World Englishes

John Speaks
University of Illinois Urbana Champaign
jspeaks2@illinois.edu

Abstract

Synonymy's inherent complexity, encompassing varying degrees of semantic overlap and context-dependent usage, poses significant challenges for computational modeling. Traditional approaches, relying on static resources like thesauruses and semantic networks, can fail to capture the full nature of synonymy, especially across World varieties of Englishes. This paper investigates and compares representations of synonymy, including manually curated resources (WordNet, Merriam Webster) and dynamically generated lists from contemporary Large Language Models (LLMs) (GPT-4o, LLaMA3.3, Deepseek-V3). With computational methods, the paper analyzes the overlap between these synonym sets using Jaccard similarity. In addition, the paper evaluates the generated synonyms within contextualized word embeddings trained on web-crawled data from Inner, Outer, and Expanding Circle countries of English. By calculating the average cosine similarity between target words and their generated synonyms within each regional embedding space, this research assesses how well different synonymy resources align with various World English varieties. The findings show significant differences in the size and content of synonym lists across methods, with LLMs behaving similarly to each other. Notably, the evaluation using regional embeddings indicates a consistent trend of higher similarity for synonyms when evaluated against embeddings trained on Inner Circle English, particularly the United Kingdom, suggesting a potential bias in both traditional and LLM-derived synonym resources towards Inner Circle norms.

1 Introduction

The concept of synonymy, the relationship between words with similar meanings, is core to human language. It allows for very specific expression, stylistic variation, and a deep understanding of semantic relationships. While seemingly straightforward, synonymy is a complex linguistic phenomenon with definitions ranging from strict identity of meaning to near-equivalence and context-dependent relationships. This complexity is a significant challenge when attempting to model it computationally, as capturing the small distinctions and contextual variations requires cleaver and nuanced approaches.

Traditional methods of representing synonymy, often relying on manually curated lists in thesauruses and dictionaries, or structured semantic networks. These approaches may not fully encapsulate the dynamic and context-dependent nature of synonymy, especially when looking at the vast linguistic diversity in the global use of English. The English language, as a globalized lingua franca, has evolved into many World English varieties, each with its own unique characteristics in vocabulary, grammar, and usage. These regional variations are often overlooked by traditional resources primarily focused on Inner Circle norms, but are crucial for a comprehensive understanding of synonymy and English in a global context.

Large language models (LLMs) have introduced new methods for representing and generating synonyms. By training on huge amounts of text data, these models can learn contextualized word embeddings that capture word meaning based on the surrounding text. This has the potential to allow LLMs to have an advantage over static representations. In this exploration, LLMs are prompted to generate lists of synonyms, providing a flexible and dynamic approach to synonym resource creation. The effectiveness of these LLM based approaches in accurately representing synonymy across World Englishes is unknown. Biases in the training data of these models may lead to an underrepresentation or misrepresentation of synonymy as it is understood and used in Outer and Expanding Circle varieties.

This paper aims to explore and compare different representations of synonymy, encompassing both

traditional resources and modern LLMs, and to evaluate how well these representations capture the nuances of synonymy across various World English varieties. The study uses computational methods to analyze synonym sets generated by different approaches and evaluates their similarity within the context of embeddings trained on data from countries representing the Inner, Outer, and Expanding Circles of English, as defined by Kachru's model (Kachru, 1985). Through the use of metrics like Jaccard similarity for set comparison and cosine similarity for synonym evaluation of closeness, this research seeks to provide a comprehensive assessment of how synonymy is represented and understood across the global varieties of English.

2 Previous Work

The very concept of synonymy presents inherent difficulties in modeling. The definition of synonymy can range from close relatedness to strict equivalence (Cruse, 1986). Different people may consider the same two words synonyms or not. Word senses also play a significant role in synonymy. One sense of a word may have different synonyms, meaning that word synonyms can be context- and culture-dependent (Ghanem et al., 2023). These factors make it difficult to create and curate lists of synonyms both computationally and manually.

Traditional approaches tend to use brute force to study synonyms and model how they related by having a set of synonyms for every sense of every word. Other approaches like Princeton WordNet are made up of manually curated lexical databases for every term (Miller, 1995). More modern LLM based approaches use contextual embeddings to take into account the surroundings of a word to disambiguate the meaning and generate synonyms. This was explored in Garcia (2021) and they found that transformer-based models using context were able to successfully disambiguate homonyms in Galician, Portuguese, English, and Spanish with context. This technology is important because it paves the way for models that can better account for the context and cultural factors outlined in Ghanem et al. (2023). At the same time, there is not yet an objective measure to determine where the threshold for a synonym, near-synonym, or unrelated word lies using any modern-day model.

A major way that context can vary for word meaning is the culture where a word or utterance appears. English, as a globalized language, has countless varieties across the world. These can be split into three main categories or circles of English as outlined in Kachru (1985): Inner Circle, Outer Circle, and Expanding Circle. The English varieties in each circle have different societal perceptions and statuses that affect how well they are represented in the world. Lovtsevich and Sokolov (2020) explored how well dictionaries represented each circle and found that Inner circle varieties of English almost completely shaped the definitions found. Outer circle and expanding circle specific definitions were rarely represented while inner circle varieties (specifically US and UK varieties) were comprehensively present.

3 Problem Definition

This paper aims to explore different representations of synonymy in language, how they compare to each other, and how well they represent different varieties of World English.

The first part of the exploration, synonymy representation comparison, is simple, as it just requires us to generate lists of synonyms and compare them with different metrics. After generating lists of synonyms with LLM prompting, API calls, or function calls, a number of metrics are used to explore the differences and identify particular patterns of interest.

The second, evaluating representation, is a little bit more complex. There is no ground truth to what is or is not a synonym, especially for a language with as much variation as English, so there is no gold standard to evaluate against. For this reason, the methods and problems explored are framed in a way to avoid having a truth to evaluate against. Instead of scoring with a 'correct' set of words, embeddings trained on different world Englishes are used to generate similarity scores for synonym pairs. This functions more as a measure of 'goodness of fit' for a set of synonyms and the embedding model for a particular region. This is the main method of experimentation for the evaluation part of the exploration.

4 Methodology

The main methodology of this exploration can be broken down into three main stages. First, lexical embeddings were trained based on web crawled data from counties across each circle of English. These embedding models would be used later as

evaluation tools for synonyms generated with other approaches. Next, tools like WordNet, Merriam Webster, and several LLMs were used to generate sets of synonyms. Finally, a variety of methods and statistical tests were used to analyze the differences in synonym sets between world Englishes.

4.1 World English Embeddings

To create the English contextual embeddings used for evaluation, English data from 13 different countries across the three circles of English were used. The data used was a web-based corpus from Dunn (2020) containing web samples of English from the 13 target countries. Two million text samples were used for each of the 13 countries in order to create a robust embedding space for each one. The countries and circles explored are outlined in table 1.

Variety	Countries
Inner Circle	United States, Canada, United Kingdom, Ireland, Australia, New Zealand
Outer Circle	India, Singapore, Nigeria, Malaysia
Expanding Circle	China, Russia, Brazil

Table 1: Varieties of English and their corresponding countries

To train the embedding models, the samples from the corpus were cleaned using the Gensim "simple_preprocess" utility which removed punctuation, normalized text to lowercase, stripped whitespace, and tokenized words while filtering out those shorter than two characters, ensuring a cleaner and more standardized dataset for training. Next, the data was fed into a Gensim skip-gram Word2Vec model for training. The code for this stage of the process is outlined in section 1 of the associated project repository[1].

4.2 Collecting Model Synonyms

Due to the varied methods of synonym storage and generation explored in this paper, different methods were used to collect the synonym data for further exploration and analysis. More direct methods were used to collect WordNet and Merriam Webster synonyms, and prompting tasks were used for the LLM approaches.

To be able to compare approaches to synonymy, a set of target words had to be selected for analysis. To make the process random and fair, the vocabulary of each country embedding model trained was intersected to create a common vocabulary to all models. 1000 words were then randomly sampled from the common vocabulary for analysis. The code for this sampling and the subsequent synonym generation is detailed in section 2 of the project code repository[2].

4.2.1 WordNet and Merriam Webster Synonyms

Collecting WordNet and Merriam Webster synonyms was a straight forward process. To get synonyms WordNet synonyms, the 'wordnet' object from the NLTK python package was used directly, giving the synonyms to the target word through a singular function call. Merriam Webster synonyms were collected using the free Merriam Webster Dictionary and Thesaurus API. Using this API, synonyms were collected for all senses of each target word and combined into a singular list. For further analysis, all of the synonyms were pre-processed in the same way as tokens were for training the embeddings.

4.2.2 LLM Synonym Generation Prompting

Three different LLMs were used for generating synonyms: GPT-4o, LLaMA3.3, and Deepseek-V3. The process was similar for all three models. An API was used to create a model request for each target word with a fixed prompt. The prompt (made up of a system and user prompt) was as follows for each model:

> System: "You are a a machine that simply functions as a thesaurus. You will be given a word and you will return a map from that word to a list of synonyms. You should include as many or as few synonyms as match the word. (E.g. given 'help' you might return 'help': ['assist', 'aid', 'support'])"
>
> User: WORD

This prompt in preliminary testing was found to be the most effective. It allows the model interpretation for how they would like to use the word and does not give any context away in the user prompt.

[1] https://github.com/JTSIV1/synonymy-representations-in-world-englishes

[2] https://github.com/JTSIV1/synonymy-representations-in-world-englishes

It also allows the model to generate as many words as is necessary for the target word. This is critical because the other tools, like the Merriam Webster thesaurus, had large variability in the number of synonyms included depending on the work and it was important for the LLMs to have that freedom too for future metrics and comparison.

I used the OpenAI API to interract with the GPT model, and the LambdaAI API for the LLaMA and Deepseek model. The cost to run this experiment was approximately $0.08 total.

4.3 Evaluating Synonym Sets

The last section of the exploration used to actually evaluate and interpret the results can be found in section 3 of the project code repository[3]. Two main methods of evaluation were used. First the synonym lists from each approach were compared to each other using the Jaccard similarity metric, and to a rudimentary approach using the nearest neighbors in each model embedding space. In addition, the rudimentary approach for each embedding model was compared to every other model in order to observe simple patterns in the embeddings through Jaccard similarity.

Next, an average similarity score was generated for each of the target words and their synonyms from each model compared to every country's embedding model. For a particular word, cosine distance was taken from the target word to each of its synonyms in the list for each model, and an average cosine distance was calculated for each word and model. These were then averaged across the 1000 target words, leaving us with an average synonym cosine distance for each method of synonym generation and county embedding model pairing.

Statistical tests were performed for each step of the analysis to determine significance of results.

5 Results

Performing all of the explorations outlined in the above steps yielded various results for each segment of the exploration. Each section has accompanying statistical tests which will not be fully covered, but more details are available on the project repository.

[3] https://github.com/JTSIV1/synonymy-representations-in-world-englishes

5.1 Preliminary Generated Synonym Observations

After generating synonyms with each method, there were a number of observed differences in the generated lists for the 1000 target words. The average synonym count per word is in table 2 for each model.

Method	Average Number of Synonyms per Word
DeepSeek-V3	7.65
GPT4o	5.19
LLaMA3.3	6.33
Merriam Webster	45.21
WordNet	5.60

Table 2: Average number of synonyms per word for each method

The LLM models and WordNet generate similar lengths of list ranging from 5.19 to 7.65 synonyms per word. The Merriam Webster thesaurus on the other hand is a large outlier with over 45 synonyms per word generated on average. Looking at the generated synonyms qualitatively, there are two main causes. First, other methods typically do not include several forms of the same root word where the Merriam Webster Thesaurus will. For example, for the target work 'plenty' GPT4o had the word 'abundance' as a synonym, while Merriam Webster had 'abundance' as well as 'superabundance' and 'abundant'. The second main reason is that the Merriam Webster thesaurus includes words that have other primary meanings, but more abstract synonym relations, where other models tend to be more direct. For example, where the primary definition of every synonym to 'plenty' generated by GPT4o was very similar to the primary definition of 'plenty', the Merriam Webster list included words like 'mass', 'bucket', and 'mountain', all of which have other primary uses.

5.2 Synonym List Jaccard Similarity

The first main area of results come from the Jaccard similarity of different sets of synonyms. Each pairing was calculated independently, and can be visualized with

5.2.1 Model to Model Similarity

The average Jaccard similarity for each model of synonymy relative to the others was calculated

across the 1000 target words. The results follow in figure 1.

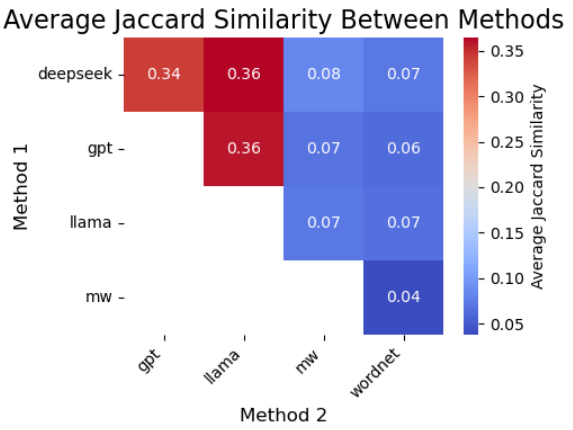

Figure 1: Model to Model Jaccard Similarity Heatmap

The LLMs have the highest Jaccard scores with each other, while every other model pairing (LLM-NonLLM and NonLLM-NonLLM) has a very low Jaccard similarity. This indicates that LLM prompts have much more overlap in how they generate synonyms, while every other pairing has little to no relation to the other methods used.

5.2.2 Model to Region Similarity

To compare each model to region representation in a rudimentary way, the number of nearest neighbors matching the number of synonyms a particular model generated for the target word was used with the list of synonyms to generate a Jaccard similarity score. This was averaged across all 1000 words for every synonym model and country pairing. The results can be seen in figure 2. The countries are grouped by circle of English, and each circle is separated by a solid black line. The order of the circles is Inner, Outer, and finally Expanding.

The results show generally that Merriam Webster and WordNet models perform worse relative to LLMs on all regions. The regional differences might show differences between specific countries, but the circle to circle difference is not significant or statistically significant.

5.2.3 Region to Region Similarity

The nearest neighbors for each region model were compared, and the similarity was plot as follows in figure 3. The regions' circle of English is separated by a solid black line where the columns and rows start with Inner Circle, then Outer Circle, and finally Expanding Circle varieties.

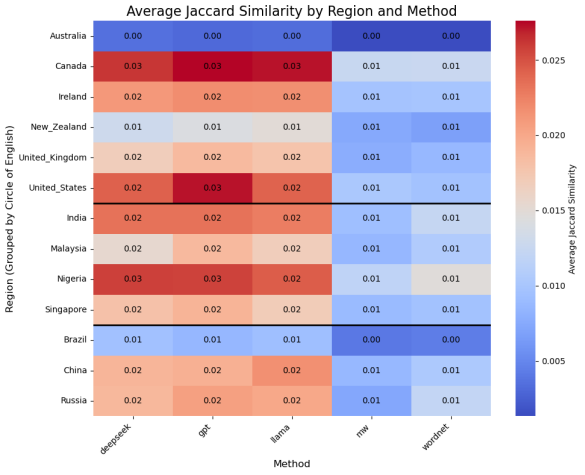

Figure 2: Model to Region Jaccard Similarity Heatmap

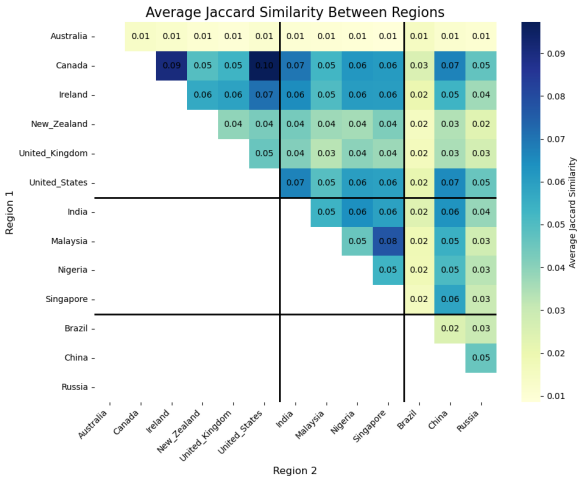

Figure 3: Region to Region Jaccard Similarity Heatmap

It is important to note that counties within each circle show significant differences to other countries in the same circle. At the same time, the nearest neighbors as a whole for all countries seem to be very dissimilar as the maximum Jaccard similarity between any two countries is 0.10. The average Jaccard similarity averaged for circle to circle comparison was not significantly different, and each circle had an average Jaccard similarity of about 0.04 with each other circle.

5.3 Evaluating Synonym Lists with Cosine Similarity

The next part of the analysis focuses on using the country embedding models as methods of ranking cosine similarity instead of using them as models to compare to. In this section, every synonym had a cosine similarity score calculated relative to its target word in every single model. The results for each country are as follows in figure 4. Dotted lines

separate the different circles of English.

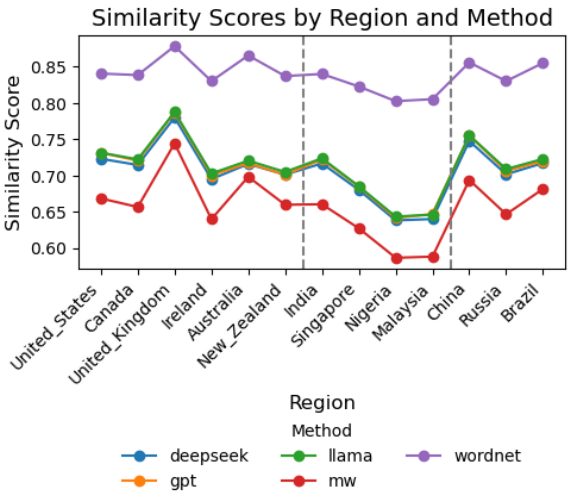

Figure 4: Synonym Cosine Similarity by Model for Each Country

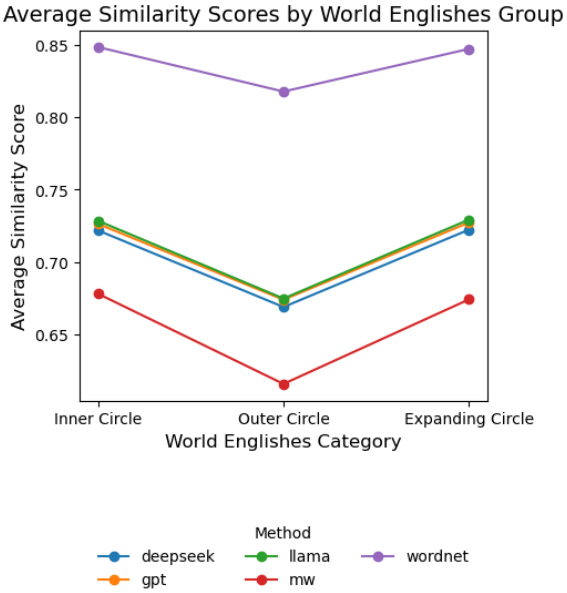

Figure 5: Synonym Cosine Similarity by Model for Each Circle of English

It is apparent that these models all follow a similar trend, with different baseline similarity levels. The WordNet (purple) line consistently has the highest cosine similarity. The LLM models perform almost identically and are not as high scoring as the WordNet model. The worst model consistently is the Merriam Webster model. The largest peak for a region is the United Kingdom. This indicates that all of the models performed best according to a United Kingdom standard of English. India and China were the best performers for the Outer and Expanding Circles respectively. The worst performers for Inner, Outer, and Expanding Circles respectively were Ireland, Nigeria, and Russia.

The analysis was aggregated by circle for further analysis in figure 5.

From these results, there is an overall trend of best performance for Inner and Expanding Circles of English and worse performance for Outer Circle varieties. Just like when separated by country, the WordNet model performs best followed by the LLMs all together, and lastly the Merriam Webster synonyms.

Next, a statistical analysis of the cosine similarity results was done. To start, the standard deviation was calculated, and the results by model for each country are shown with error bars in figure 6.

There is significant overlap in the error bars for every model. The model with the most defined differences is the Merriam Webster Thesaurus model. The significant overlap was a cause for concern, so to further test the significance an ANOVA test was performed to determine if country was a significant predictor in the resulting similarity. The test was definitive that country variation was significant. The same test was performed to see if the differences by circle of English were significant, and the ANOVA also showed a definitive relationship meaning that the difference in circle is a significant predicting factor in cosine similarity of synonyms.

For more testing, a pairwise Tukey test was performed to see what countries were or were not significantly different from one another. The test found that all country pairings were significant apart from two (97%) which were Canada and India as well as Australia and Brazil. These results mean that only those two pairings of countries do not have statistically significant similarity in terms of average synonym cosine similarity score. The same Tukey test was done, but grouped by circle of English, and found that all circles of English are statistically significant from one another.

ANOVA tests were also done for each synonym model instead of just with the aggregated results. The ANOVA tests showed that country played a statistically significant role for each model. A Tukey test was done for each model of synonymy as well, and while some country pairings were not significant for certain models, the overwhelming majority were significant in every model.

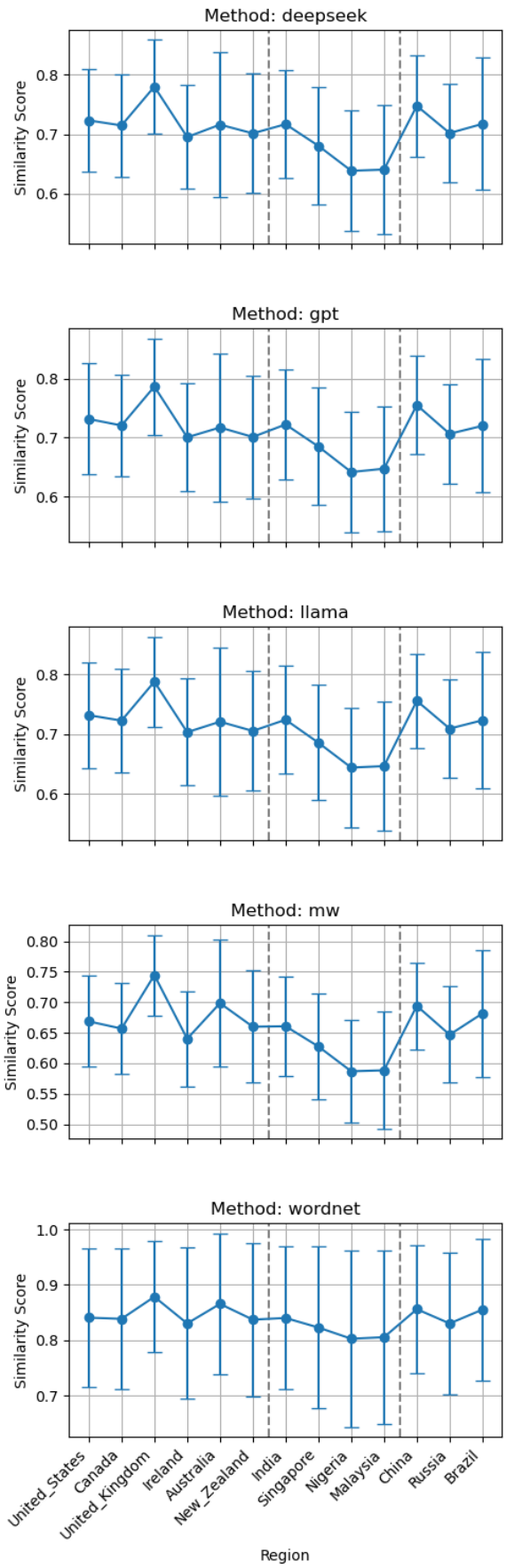

Figure 6: Cosine Similarity with Error Bars for Each Method and Country

6 Discussion

The findings of this study offer several key insights into the representation and evaluation of synonymy across different varieties of English. First, the stark contrast in the average number of synonyms generated by different methods (Table 2) underscores the inherent variability in how synonymy is conceptualized and curated. The inclusion of additional words in the Merriam Webster synonym lists indicate that Merriam Webster is less conservative than the other methods observed when it comes to what is considered a synonym, and it has a more general definition. The qualitative results showed that while several word forms for each word contributed to the inflated word count, more distantly related words being included also contributes to the higher synonym count.

When looking at the model to model Jaccard similarity, it was interesting to see that the models all performed fairly differently, but the LLMs all had the highest overlap. This indicates that, while all trained differently, the structure of each LLM responded to the prompting task similarly to the other LLMs. In addition the Merriam Webster and WordNet approaches were extremely dissimilar to all other approaches in the Jaccard similarity of synonyms.

The other Jaccard similarity tests like the model to region and region to region comparisons were not as insightful. While there were patterns present and the results were statistically significant, the results were not practically significant. For both of these segments of exploration, the maximum Jaccard similarity was under 0.1, indicating that pretty much every pairing was dissimilar. A reason for this could be due to flaws in this approach for embedding models. Nearest neighbors in an embedding model are typically words used in the same context, not necessarily synonyms. When using this as a rudimentary way to emulate synonyms, issues arise because synonyms are similar in the models but so are many other irrelevant words like antonyms. It is also possible that the data was not a good representation of English and could have unclean data, but a qualitative analysis of some of the embeddings and their neighbors did not find any issues.

Looking at the tests related to cosine similarity, with each embedding space as the evaluation metric, not a point of comparison, the results were more significant and impactful. Looking at model

variation by country, the United Kingdom had better model performance than other countries across all models. This is consistent with the societal perception of English in the United Kingdom as the most proper form. This perception translates into most models of synonymy best matching English from the United Kingdom. The results by circle of English are also consistent with theory. Kachru (1985) theorizes that Inner Circle varieties of English are more accepted than Outer Circle varieties which is supported by the results. Additionally, it is thought that Expanding Circle varieties of English pattern more like Inner Circle varieties because they have not yet become distinct, and learners of English in these countries learn from Inner Circle speakers.

7 Limitations

This study, while providing insights into the representation and evaluation of synonymy across World Englishes, has a number of limitations to consider.

The study relied on a Word2Vec skip-gram model for training the regional word embeddings. While Word2Vec is a widely used and effective method for capturing semantic relationships, more advanced contextual embedding models, such as Transformer-based architectures (e.g., BERT, RoBERTa), might capture more nuanced semantic differences across varieties of English. Future research could explore the use of these more sophisticated models for evaluation in a similar study.

Furthermore, the generation of synonyms from Large Language Models was based on a specific prompt designed to elicit thesaurus-like responses. Different prompting strategies or parameters could potentially yield different sets of synonyms. This paper did not look at the sensitivity of LLM-generated synonyms to variations in prompting, but future explorations could look into how tweaks to prompting can be used to increase task performance.

The evaluation of synonymy relied primarily on Jaccard similarity for set overlap and cosine similarity within the embedding spaces. While these are standard metrics in this field of analysis, they may not fully capture the complexities of synonymy. Jaccard similarity, for example, is sensitive to the size of the synonym sets, and cosine similarity, can capture semantic relatedness but, does not directly measure synonymy and can be influenced by other linguistic relationships.

Finally, the selection of countries to represent the Inner, Outer, and Expanding Circles of English, based on Kachru's model, provides a useful framework but is a simplification of the complex landscape of English around the world. There is considerable variation within each circle (and even within each country), and the chosen countries might not fully represent the diversity within those circles.

8 Conclusion

This paper has explored and compared various representations of synonymy, including those derived from more traditional resources like WordNet and Merriam Webster, as well as those generated by new large language models such as GPT-4o, LLaMA3.3, and Deepseek-V3. The analysis revealed significant differences in the breadth and nature of synonym lists produced by these methods, with LLMs showing more overlap between themselves compared to the traditional resources. In addition, when evaluating these synonym sets against embeddings trained on different World Englishes, it was observed that the United Kingdom's variety consistently yielded the highest cosine similarity scores across all synonym generation models, aligning with existing linguistic hierarchies and perceptions. Notably, the LLMs had the same pattern of performance on each country as the traditional methods. This suggests that the training data of these models also contains the same bias towards Inner Circle varieties of English and English in the United Kingdom specifically. While the training data is largely unknown for the models, the results indicate that the corpora used are likely not representative of the way people around the world speak English. The statistical significance of the differences in similarity across countries and circles of English further reinforces the context-dependent nature of synonymy. While the simple approach of using nearest neighbors in embedding spaces for cross-regional comparison was not as insightful, the evaluation with cosine similarity provided a good look into how well different synonymy resources align with semantic representations learned from diverse English-speaking regions. These findings highlight the complexities of defining and modeling synonymy in a globalized language like English and continue to reinforce the idea that legitimate varieties of Outer Circle English are not accepted as proper English and are underrepresented.

References

D.A. Cruse. 1986. *Lexical Semantics*, volume 1. Cambridge University Press & Assessment.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. 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)*, pages 3625–3640, Online. Association for Computational Linguistics.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching Arabic synonyms. In *Proceedings of the 12th Global Wordnet Conference*, pages 274–283, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Braj B. Kachru. 1985. Standards, codification, and sociolinguistic realism: The english language in the outer circle. In Randolph Quirk and Henry G. Widdowson, editors, *English in the World: Teaching and Learning the Language and Literatures*, pages 11–30. Cambridge University Press.

G. N. Lovtsevich and A. A. Sokolov. 2020. World englishes and learner lexicography: View from the expanding circle. *Russian Journal of Linguistics*, 24(3):703–721.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.