Dialectal Semantics: Measuring Variation Between Indian and American English

Michelle Yang
McGill University, Linguistics
michelle.yang5@mail.mcgill.ca

Robert Dow
McGill University, Cognitive Science
robert.dow@mail.mcgill.ca

Abstract

There exist two prominent schools of thought when it comes to the universality of word meaning across language. Universalists propose that semantic meanings are the same cross-linguistically, whereas relativists propose that they can vary. While many prior studies have aimed to investigate this debate (Youn, Sutton, Smith, Moore, Wilkins, Maddieson, Croft, and Bhattacharya, 2016; Berlin and Kay, 1991), word meanings, unlike other linguistic features, are difficult to study directly as they are not observable. However, the development of distributional semantic models (see Jones and Mewhort (2007); Landauer and Dumais (1997); Mikolov, Chen, Corrado, and Dean (2013)) have allowed us to investigate this question more deeply. Recent findings suggest that languages do differ cross-linguistically, such that those that are less culturally, geographically, and linguistically related have less aligned word meanings (Thompson, Roberts, and Lupyan, 2020). Further research found that bilingual populations think along the semantic representations from their L1, even when speaking in the L2 (Lewis, Cahill, Madnani, and Evans, in press). This study seeks to expand these findings by investigating English as a lingua franca – specifically English dialects in Kachru's outer circle of World English (Kachru, 1992). English, as spoken in places such as India, has been "institutionalized as an additional language" (Kachru, 1992) to the various languages already spoken in the area. As such, these contact dialects of English have acquired stable and distinct characteristics in their linguistic features. While many studies have investigated the phonetic, syntactic, and discoursal features of contact Englishes, relatively few studies have investigated their semantic features. This study aims to explore whether dialects of English exhibit differences in their representation of word meanings, and more broadly, whether cross-linguistic contact influences the semantic space of a language.

1 Introduction

The extent to which the meanings of words vary between languages is a question that is debated and investigated. A universalist view would argue there exists an innate structure of language for all humans; therefore, all human language draws from a pre-existing conceptual space and any cross-linguistic differences do not affect cognition (Regier, Kay, Gilbert, and Ivry, 2010). The relativist view, on the other hand, suggests that languages are free to vary (Whorf, 1956), and differences between languages reflect different solutions to categorizing thought (Thompson et al., 2020). The main point on which universalist and relativist perspectives differ is on common, everyday, meanings. Universalists predict that common concepts, such as the self, animals, emotions, objects, are available to all, regardless of language. As such, any differences across languages should be random and unpredictable. Relativists argue the opposite, that even seemingly "simple" concepts may not align cross-linguistically. Importantly, the relativist view anticipates that these alignments should be predictable from cultural, geographic, and linguistic features (Thompson et al., 2020).

With the use of distributional semantic models like word2vec (Mikolov, Chen, Corrado, and Dean, 2013), researchers have been able to study the semantic universality vs relativity with more objective approaches. Recent work has been done to quantify how words semantically vary between languages by utilizing these models. Thompson et al. (2020) demonstrated that semantic neighborhoods between language and words' semantic similarities vary by a degree of their historical relationships, geographic proximity, and cultural similarity. Lewis et al. (in press) builds on this work by documenting specific local and global similarities between languages. Additionally, this work provides some of the first large-scale investigations into semantic differences between languages.

While this previous work provides insight into the way that word meanings vary across languages, it is still unclear how much word meanings may vary within languages. As a result of English's unparalleled spread across the globe, there is the rise of what is known as the World Englishes. Indian English is one such variety. World Englishes co-exist with local languages and as such, are influenced by the additional languages these speakers have; while many studies have described the distinct phonological, lexical, and grammatical features of these varieties of English (Kirkpatrick, 2014), there has been less study on their semantic features.

To further investigate this, we present a large-scale, cross-dialectal analysis of the semantic space of Indian English speakers. Our analysis seeks to understand the extent to which contact with other languages, as well as the distinct cultural environments in which English is spoken affects word meanings. To do so, we will use word2vec word embeddings trained on Indian English and compare that with word2vec word embeddings trained on American English (Continental United States). This analysis will allow us to investigate how much certain semantic domains (e.g. animals, body, numbers) align across dialects. The universalist view would predict high alignment across concrete concepts, especially within one language; in comparison, the relativist view would predict that cultural differences should modulate alignment. This analysis seeks to provide new insights into the universality of language and how semantic variation reflects the culture, history, and geography of speakers.

2 Background

To motivate our analysis of Indian English as a dialect of English, we use Kachru's Circles of World Englishes model (Kachru, 1992). These are concentric circles known as the inner, outer, and expanding circles. The inner circle are countries in which English is the first language of a majority of the speakers (e.g. Great Britain, United States). The outer circle are mostly postcolonial nations, in which English occupies an official role (e.g. India, Singapore). Lastly, the expanding circle consists of countries where English is used as a foreign language, and occupies no official role (e.g. China, Egypt). While there are many stages that the outer circle World Englishes go through (Schneider, 2007), the last stage is known as the differentiation stage. In this stage, group identities become more important, and as result, are reflected in dialectal differences (Schneider, 2007). Our study will focus on Indian English as it is one such variety that researchers have argued has reached the final stage (Kirkpatrick, 2014; Balasubramanian, 2009). While there has been some criticism of Kachru's model, it is still largely considered the most influential model of World Englishes and any critiques of its shortcomings are out of the scope of this study.

Indian English, while mutually intelligible with other varieties of English (i.e. Canadian English, Scottish English) has its own distinctive grammatical features, with a vocabulary that reflects the local culture (Kirkpatrick, 2014). However, in linguistic studies of Indian English, typically only the phonological, syntactic, and lexical features are described. Varieties, such as Indian English, are often considered to have undergone substrate transfer – influence from features of the first language of the speakers onto features of English (Sailaja, 2012). As such, some of the unique features of such varieties can be explained by their presence in the first language. For Indian English, substrate transfer is attributed to certain phonological, lexical, and syntactical features. Importantly, the uniqueness of Indian English is not only due to substrate influence, but also the cultural societal requirements (Sailaja, 2012). Recently, researchers have conducted more investigations into the pragmatic features of World English (Kirkpatrick, 2014). In a 2010 study of "Persian" English, researchers found that different Englishes express the cultural notions of their speakers (Sharifian, 2020). With evidence of substrate influence in other linguistic features of Indian English, as well as general findings that varieties of English express certain concepts in unique ways, it seems pertinent to now investigate the unique semantics of Indian English.

It is difficult to study word meanings in natural languages. After all, how can we say for sure that two (or more) words have the same meaning? Even close synonyms are slightly different in meaning (e.g. love vs adore). One reason why the semantic structure of languages is difficult to quantify is that word meanings are not directly observable. Prior studies have attempted to do so by asking speakers to rate word similarity of a word and its synonyms (Magué, 2006). However, using such methods is time consuming, difficult, and subjective. With the recent development of neutral network approaches to word embeddings, it is now possible to more quantitatively describe word meanings. This class of models are known as distributional semantic models (DSMs). DSMs are based on the distributional hypothesis, which states that "you shall know a word from the company it keeps" (Firth, 1957). In contemporary instantiations, DSMs are systematically represented using vector spaces (Lenci, 2018). These models are capable of deriving semantic representations of words based on their distribution in large text corpora. The end result is vector representations of word meanings, also known as word embeddings. In this framework, distance between vectors corresponds to word similarity, such that closer words have more similar meanings. Word embeddings have been shown to be correlated with human judgements of word similarity and can account for a variety of lexical semantic data (Hill, Reichart, and Korhonen, 2015; Johns, 2021). Furthermore, word embedding representations are able to capture the range of contexts in which a word is used and the relative frequency of those contexts (Thompson et al., 2020). As such, it is possible to compare words across multiple contexts – in our case, Indian English vs American English.

In this study, we will create static word embeddings for both the American English and Indian English corpora using the skip-gram word2vec (Mikolov et al., 2013) algorithm. The skip-gram algorithm modifies the projection layer in a neural network architecture such that it predicts the surrounding words of a target word given an input of encoded text. Compared to its continuous bag of words (CBOW) counterpart which performs better on syntactic tasks, the skip-gram model performs better on semantic tasks. Thus, this algorithm is ideal for assessing the latent semantic differences between Indian and American English.

Prior work has investigated cross-linguistic variation in the structure of semantic space and in English L2 bilinguals (see Thompson et al. (2020); Lewis et al. (in press)) using vector representations. These studies found that there is substantial variation across languages in the structure of their semantic spaces, but that this variation is predictable. Languages that are culturally, geographically, and linguistically similar show greater similarity in their semantic structure as well. Thompson et al. (2020) found the highest degree of semantic similarity in domains with high internal structure – numbers, temporal terms, familial terms – suggesting that these meanings are derived from a universal cognitive and perceptual base. Domains like common objects, actions, and natural objects only had intermediate similarity cross-linguistically. Moreover, they found that across all domains, similarity was predicted by non-linguistic measures like historical, geographical, and cultural relatedness. Lewis et al. (in press) find similar results in a study of L2 English writing. They find that even highly-skilled L2 bilinguals exhibit semantic variation that diverges from L1 English. Most importantly, they found that this variation in L2 speakers is modulated by the features of the native language. In addition, they find that languages tend to be locally similar, but globally varied. That is, languages are similar in how word meanings cluster to each other, but those clusters themselves show variation across the semantic space as a whole. Overall, the results from both studies suggest that culture has influence on the semantic structure of languages and that there are some universal patterns that the world's languages follow.

3 Baseline

To maximize comparability with prior research in this field, we use Thompson et al. (2020) as a baseline. As such, we use their 21 domains to compare Indian and American English (listed in order of most to least aligned in Thompson et al. (2020)): Quantity, Time, Kinship, Function Words, Animals, Sense perception, The physical world, Cognition, Food and Drink, Possession, Spatial relations, Speech and language, The body, Social and political relations, Emotions and Values, Clothing and grooming, Agriculture and vegetation, Modern world, Motion, Basic actions and technology, and The house.

In addition, we use their semantic alignment algorithm (described in more detail in 5. Methodology) to allow for direct comparison with their cross-linguistic analysis. This algorithm builds second-order word vectors that compare the semantic similarities of word pairs across corpora. Under the distributional hypothesis, words that occur in similar contexts should have similar meanings to each other (Firth, 1957). As such, by comparing the vector representations of words and concepts derived from our models, we are able to compare their meanings.

4 Dataset

To train our models, we used two different corpora. The first was IndicCorp (Kakwani et al., 2020), an Indian news text corpus, and the other was a corpus of U.S. news text called the North American News Text Corpus (NANTeC) (Graff, David, 1995). We chose corpora from the same genre to ensure comparability across models. The underlying assumption behind these corpus selections is that they contain text about similar topics, and thus, contain similar words in similar contexts, reducing the possibility that our results arise from genre-specific differences.

NANTeC

The North American News Text Corpus (NANTeC) (Graff, David, 1995) is composed of English text gathered from U.S. news outlets between 1994 and 1997. The corpus totals 350 million words as broken down in Table 1. After pre-processing the text (see 5. Methodology), there were 46.6 million remaining words that were used in this study.

News Outlet	Date Range	Size (words)
Los Angeles Times/ Washington Post Service	05/94-08/97	52 mil
New York Times News	07/94-12/96	173 mil
Reuters News Service	04/94-12/96	85 mil
Wall Street Journal	07/94-12/96	40 mil

Table 1: Breakdown of sources used in the NANTeC dataset

Text in the NANTeC is formatted using TIPSTER-style SGML markup that we used to collect article content. An example snippet follows:

"<p> By conventional wisdom, there are certain things you simply don't do, right? <p> You don't drink on an empty stomach. You don't spit into the wind and, of course, you never escort the bride's father to the bachelor party. <p>"

IndicCorp

IndicCorp (Kakwani et al., 2020) is a corpus of Indian languages consisting primarily of news, magazines, and books from thousands of sources. In this study we used the English subset of this corpus. This subset contains 3.49 million articles which comprise 54.3 million sentences or 1.22 billion words. Of these 1.22 billion words, 350 million pre-processed words were used in this study. In this corpus, every line is a sentence from an article. An example line from this corpus follows:

"If there is one visual metaphor that explains this survey, it is that of a vast battlefield where a lost war is being fought."

5 Methodology

The NANTeC and IndicCorp corpora were preprocessed to have the same structure before being used to build semantic word representations. First, titles and meta-data such as copyright information were removed, leaving only the article body. Next, punctuation and non-ASCII characters were removed, and words were lower-cased. Finally, the resulting data were divided into sentences and tokenized.

Next, the tokenized sentences of each corpus were used to create word embeddings via the Word2Vec skip-gram model (Mikolov et al., 2013). This model uses deep learning to learn word embeddings by masking words in a sentence and trying to predict the masked word based on its surrounding words. This training procedure was carried out using the Gensim library (Řehůřek and Sojka, 2010). The implementation used Gensim's default hyper-parameters with the exception of vector size which was set to 200.

Using the trained word embeddings from each corpus, we computed the semantic alignments of words between the two English dialects following the procedure outlined by Thompson et al. (2020). To calculate the semantic alignment of a word w, we first found the $k = 100$ closest semantic neighbors to w from one dialect D_1 that existed in the other dialect D_2. For each dialect, the vector cosine similarity between w and the other k words was computed and the Pearson correlation value was calculated between these two sets of word similarities. This represents the directional semantic alignment from $D_1 \rightarrow D_2$. We can see an example of this in Figure 1 where the most similar words to "heat" in Indian English are "sweltering" and "humid". While these words have a similarity of 0.67 and 0.66 to the word "heat" in Indian English, in American English, these words are less related, being associated with similarity values of 0.51 and 0.47. This result indicates low directional semantic alignment from Indian to American English. Continuing with this implementation, the directional semantic alignment was subsequently computed from $D_2 \rightarrow D_1$. The average of these two results was taken to represent the semantic alignment of word w between both dialects.

Lastly, we computed the alignment of semantic domains across dialects. Semantic domains such as "time" or "the body" consist of sets of related words around a specific concept that commonly appear across languages and cultures. Using the list of domains outlined in Thompson et al. (2020) and derived from the NorthEuraLex dataset (Dellert, Daneyko, Münch, Ladygina, Buch, Clarius, Grigorjew, Balabel, Boga, Baysarova et al., 2020), we ended up with 885 total words across 21 domains which appeared in both English dialects. The alignment of each domain was calculated by finding the averaging semantic alignment of all words in its domain. The alignment value of a given word, domain, or dialect is henceforth referred to as a.

5.1 Model Selection

We chose the Word2Vec skip-gram algorithm because it has been shown to best account for semantic meanings across large text-corpora (Mikolov et al., 2013). Furthermore, the Word2Vec model has been shown to align with human judgements, as well as variance in various behavioural measures of lexical access (Hollis and Westbury, 2016). The skip-gram algorithm organizes meaning in a way that is comparable to prior findings in psycholinguistic research (Hollis and Westbury, 2016). As such, it is suited for our investigation into the organization of the lexical semantic space.

6 Results

Alignment of Indian English and American English varied by domain, as shown in Figure 2. The top three domains with highest alignment are Quantity, Kinship, and Time, which mirrors the findings of Thompson et al. (2020). Interestingly, in our analysis the three least aligned domains are Food and Drink, Animals, and Agriculture and Vegetation. In comparison, the least aligned domains as shown in Thompson et al. (2020) are Motion, Basic actions and Technology, and the House. See Table 2 for a full comparison of our findings.

7 Discussion

For Indian English and American English, the most aligned domain is Quantity (alignment, $a = 0.61$). This is unsurprising given the fact that both dialects of English use the base-ten number system and the same orthographic forms. Similar to the reason proposed by Thompson et al. (2020), it seems that there is a cognitive constraint that arises from the base-ten system that results in similar representations of quantities across groups of speakers. Adjacent research has investigated the link between counting systems and cognition, finding that languages that have smaller counting systems also have greater difficulty expressing large numerical values (Beller and Bender, 2008), reflecting the link between language and cognition.

While we expected similarities to the findings of Thompson et al. (2020), we did not expect complete alignment as we were comparing cross-dialectally between two dialects, as opposed to cross-linguistically across 41 languages. However for certain domains, the differences were greater than expected. In particular, we found that the semantic domain of Animals is one of the least aligned ($a = 0.16$), whereas Thompson et al. (2020) found that it was one of the top 5 most aligned domains. While this is unexpected, we attribute this to genre-specific limitations. Our training materials were limited mostly to news text. As such, it is likely

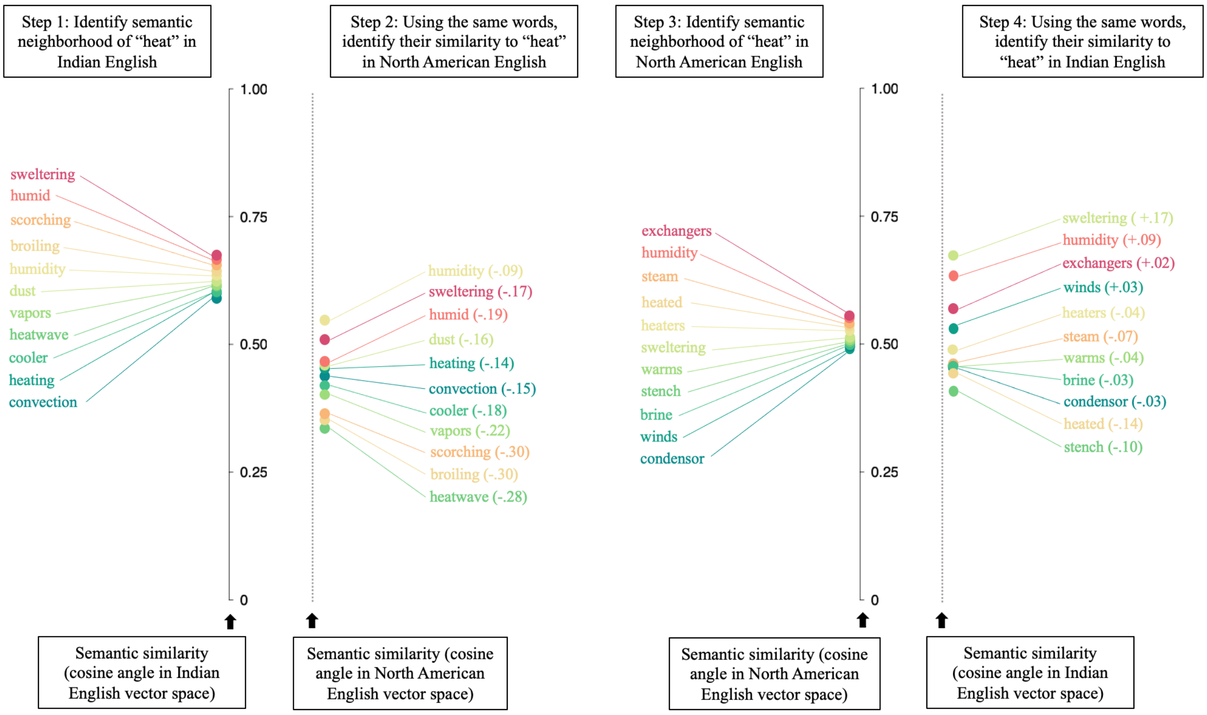
Step 1: Identify semantic neighborhood of “heat” in Indian English
Step 2: Using the same words, identify their similarity to “heat” in North American English
Step 3: Identify semantic neighborhood of “heat” in North American English
Step 4: Using the same words, identify their similarity to “heat” in Indian English
1.00
0.75
0.50
0.25
0
sweltering
humid
scorching
broiling
humidity
dust
vapors
heatwave
cooler
heating
convection
humidity (-.09)
sweltering (-.17)
humid (-.19)
dust (-.16)
heating (-.14)
convection (-.15)
cooler (-.18)
vapors (-.22)
scorching (-.30)
broiling (-.30)
heatwave (-.28)
exchangers
humidity
steam
heated
heaters
sweltering
warms
stench
brine
winds
condensor
sweltering (+.17)
humidity (+.09)
exchangers (+.02)
winds (+.03)
heaters (-.04)
steam (-.07)
warms (-.04)
brine (-.03)
condensor (-.03)
heated (-.14)
stench (-.10)
Semantic similarity (cosine angle in Indian English vector space)
Semantic similarity (cosine angle in North American English vector space)
Semantic similarity (cosine angle in North American English vector space)
Semantic similarity (cosine angle in Indian English vector space)

Figure 1: Semantic alignment of the word "heat" between American and Indian Dialects of English. This figure outlines the steps of calculating semantic alignment. This figure is adapted from Thompson et al. (2020).

that animal-domain terms, such as "spider, flock, pasture" did not appear in significant quantities across the corpora, resulting in weak word representations.

Thompson et al. (2020) calculated a baseline of concepts across two English corpora and found the average alignment was $a = 0.53$, suggesting that there is some within-language differences. In comparison, we found that the average alignment across American and Indian English is $a = 0.32$. This alignment is both less than 50% alignment and smaller than the findings of Thompson et al. (2020), suggesting that there does exist some differences in the semantic spaces of American and Indian English that cannot be simply ascribed to corpora differences.

More of interest to our research question is the fact that there do indeed exist semantic differences between Indian and American English. Under the Universalist model, we would expect few differences cross-linguistically as different languages should organize the world in similar ways, reflecting similar cognitive constraints. We would then expect even fewer, if not zero, differences within dialects of the same language. Under the Relativist model, different languages are expected to have different structures and ways of organizing concepts, which could reflect cultural differences; however, little is said about what is expected cross-dialectally.

We propose that our findings fit neither the Relativist nor Universalist models. Since we found that Indian and American English do not fully align, our results do not support the idea that all speakers map the same words onto the same concepts; however, like (Thompson et al., 2020) we also find that domains with high internal structure like Quantity, Kinship, and Time, are more aligned. This suggests that concepts like the base-ten number system, the 12-month Gregorian calendar, and familial relationships does constrain the meaning of words in those domains across cultures that share in these systems. These concepts seem to have some basis in the natural world – for example, the Gregorian calendar is a solar calendar that is based on the Earth’s rotation around the sun and is the most widely used around the world (Reingold and Dershowitz, 2018). There is a universality in the Earth’s rotation and also its rotation around the sun, which suggests that there may be a common perceptual mapping onto these concepts from the natural world. Furthermore, since our findings are similar to that of (Thompson et al., 2020), this suggests that for certain concepts there do exist cross-linguistic and cross-cultural universalities.

We found using a distributional semantic approach, that there are marked differences in the semantic spaces of Indian and American English and that these differences vary by domain. While it is difficult to know exactly what accounts for this difference in semantic alignment, we propose that it is from a mixture of syntactic, discoursal, and/or cultural differences. Indian English has a number of syntactic and discourse-pragmatic features that are different from American English, particularly in spoken language (Lange, 2012). These differences in word-usage and word co-occurences may have an effect on the word representations learned by the Word2Vec skipgram algorithm (Mikolov et al., 2013).

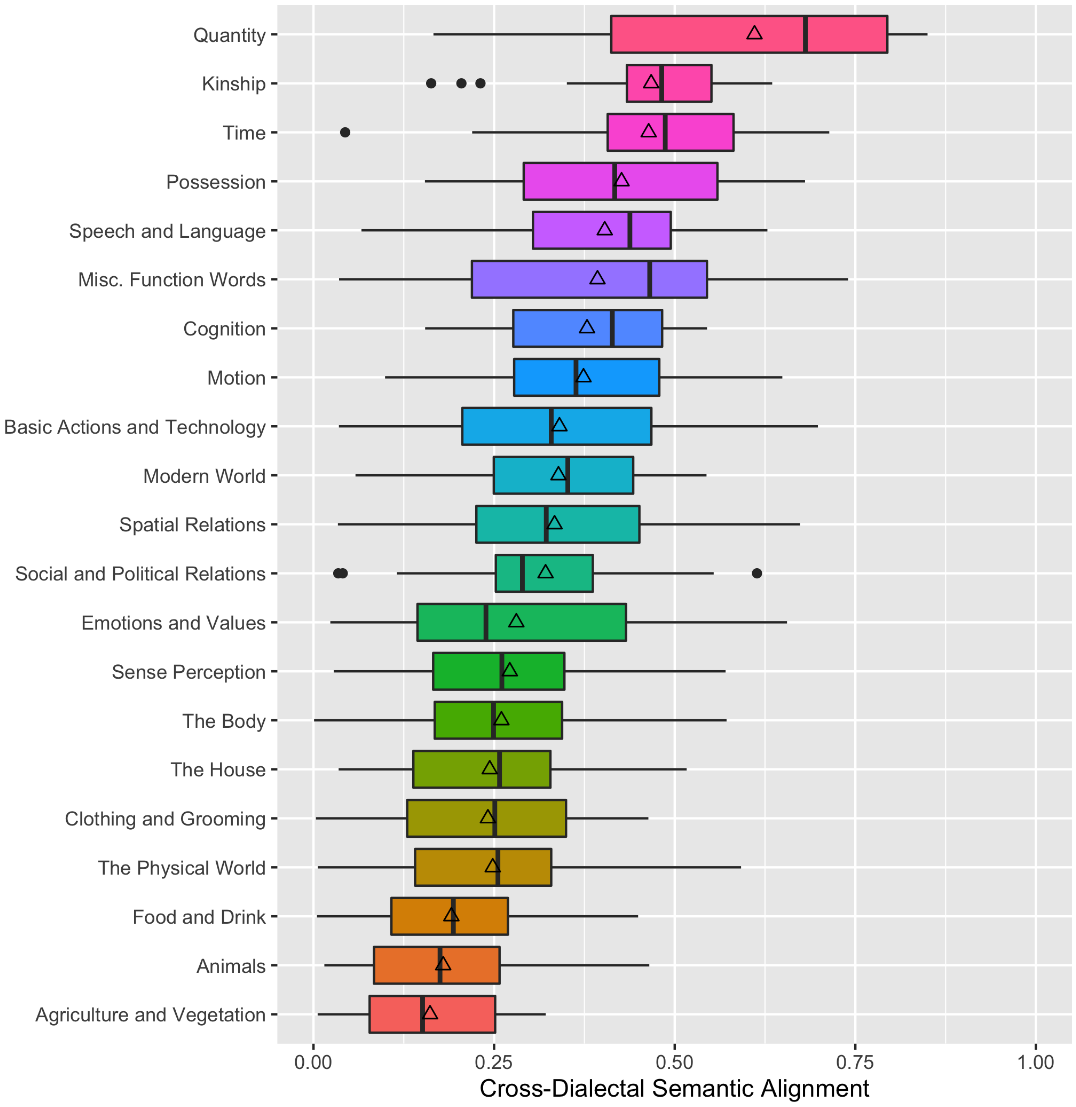

Figure 2: Cross-Dialectal Semantic Alignment Across 21 Domains as ranked by mean alignment across domains. The triangle indicates mean alignment across all words in a given domain for IndicCorp and NANTeC. The black line in the box plots represents the median, and the limits of the box show the Interquartile Range (IQR), which represents the 25th-75th percentile. The whiskers show 1.5 times the IQR. This plot is adapted from Thompson et al. (2020).

This cross-linguistic contact, resulting in substrate transfer, does have some influence on the semantic space of a language. An alternative (or perhaps concurrent) explanation is that Indian English reflects slightly different solutions to categorizing the world, actions, and thoughts. This suggests that even for a single language, speakers from different cultural backgrounds and geographical areas may organize their language differently from each other.

7.1 Limitations

Our analysis is limited in three significant ways. First, the corpora sizes differ significantly. Since we used a distributional semantic model, the quantity of training data plays a significant role in the quality of the word vector representations. We attempted to account for this initially by randomly selecting a subset of IndicCorp to match the size of NANTeC as a way to rule out any differences in vector representations due to training size; however, after data cleaning, NANTeC was still smaller than IndicCorp. While NANTeC still generated sensible and usable word vectors, it would have been preferable to have comparable corpus sizes. (For comparison, the Google News Pretrained Model contains 100 billion words (Řehůřek and Sojka, 2010)). While there are larger corpora of English news text, in our initial design we prioritized controlling for geographic area/dialect variety. Future research would greatly benefit from larger training corpora of both Indian and American English.

Another limitation is that NANTeC (Graff, David, 1995) only contains text data from the late 90s. In com-

Our Findings	Thompson et al. 2020
Quantity	Quantity
Kinship	Time
Time	Kinship
Possession	Misc. Function Words
Speech and Language	Animals
Misc. Function Words	Sense Perception
Cognition	The Physical World
Motion	Cognition
Basic Actions and Technology	Food and Drink
Modern World	Possession
Spatial Relations	Spatial Relations
Social and Political Relations	Speech and Language
Emotions and Values	The Body
Sense Perception	Social and Political Relations
The Body	Emotions and Values
The House	Clothing and Grooming
Clothing and Grooming	Agriculture and Vegetation
The Physical World	Modern World
Food and Drink	Motion
Animals	Basic Actions and Technology
Agriculture and Vegetation	The House

Table 2: Comparison of our findings to (Thompson et al., 2020). Domains are listed in descending order in terms of their semantic alignment where our results represent alignment between Indian and American English, and Thompson's results are the average alignments across unique pairs of 41 languages.

parison, IndicCorp (Kakwani et al., 2020) is scraped from various internet sources – mostly news, magazines, and books – meaning that it contains more contemporary text. This may influence certain word representations as certain words have changed in and gained meaning over time (e.g. cloud now refers to both the weather phenomenon and computational resources). Due to the way the skip-gram algorithm learns word representations, these polysemous words would all have one word vector representation. As such, future research should account for these polysemous meanings either by choosing corpora from similar time periods, or by factoring this into the calculation for semantic alignment using some other method.

In addition, because we wanted to control for genre (and due to sheer availability) we used two news corpora. However, news corpora are not necessarily ideal for studying dialectal variation. From a sociolinguistic perspective, the ideal data is spontaneous and casual. Written text, like news sources, are heavily edited and carefully constructed. Furthermore, "nonstandard" constructions are often subjected to various prescriptivist rules. These factors may reduce the occurrence of dialect-specific word co-occurrences, which would influence semantic alignment results. Future research should try to train semantic models on casual and spontaneous corpora, as these would be the most representative of variation across dialects.

In this analysis, we were unable to account for why there are differences in the semantic alignment of Indian and American English. While we could hypothesize that this is due to influence from the structure of Indian English and/or cultural differences, it is impossible to have a deeper understanding of how these differences arise with our current methodology. One way to do so would be to train another model on a language spoken in India, such as Hindi or Telugu, to have a three-way comparison of semantic alignment. This way, it would be possible to see if Indian English more closely aligns with its dialectal counterpart or with another language that is spoken in the same geographical area and by those of the same cultural background. If Indian English were to more closely align with another language spoken in India, this could suggest that cultural similarity of speakers plays a role in speakers' organization of their lexical semantic space. Future research would greatly benefit from exploring this direction more.

8 Conclusion

Our findings provide support for the Relativist view, such that languages do not necessarily reflect innate concepts, but may vary by speaker and by culture. However, our findings also support the Universalist view. The domains of Quantity, Time, and Kinship seem to reflect some common cognitive organization of these categories.

Our work adds to the growing literature that seeks to understand how languages and speakers organize meaning in their minds. As more corpora, models, and methods are developed, quantifying the semantic differences between the world's languages and dialects will only get better. We hope that further research in this area will allow us to understand the effects of distinct cultural and linguistic environments on the development of word meanings. While our findings cannot rule out either a Universalist or Relativist view, we see this as a step in the right direction towards increasing our understanding of semantic variation.

9 Contributions

The authors contributed equally to this project.

10 Acknowledgements

We would like to thank Esteban Hernandez-Rivera for his helpful comments and feedback on this paper.

References

Chandrika Balasubramanian. 2009. *Register Variation in Indian English*. John Benjamins Publishing Company.

Sieghard Beller and Andrea Bender. 2008. The limits of counting: Numerical cognition between evolution and culture. *Science*, 319(5860):213–215.

Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution*. Univ of California Press.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, et al. 2020. Northeuralex: A wide-coverage lexical database of northern eurasia. *Language resources and evaluation*, 54(1):273–301.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Graff, David. 1995. North american news text corpus.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin &amp Review*, 23(6):1744–1756.

Brendan T Johns. 2021. Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*, 131:101441.

Michael N Jones and Douglas JK Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, 114(1):1.

Braj B Kachru. 1992. World englishes: Approaches, issues and resources. *Language teaching*, 25(1):1–14.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Andy Kirkpatrick. 2014. *World Englishes*. Routledge.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Claudia Lange. 2012. *Syntax of Spoken Indian English*. John Benjamins Publishing Company.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Molly Lewis, Aoife Cahill, Nitin Madnani, and James Evans. in press. Local similarity and global variability characterize the semantic space of human languages.

Jean-Philippe Magué. 2006. Semantic changes in apparent time. In *Annual Meeting of the Berkeley Linguistics Society*, volume 32, pages 227–235.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Terry Regier, Paul Kay, Aubrey L Gilbert, and Richard B Ivry. 2010. Which side are you on, anyway? *Words and the Mind*, page 165.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Edward M. Reingold and Nachum Dershowitz. 2018. *Calendrical Calculations: The Ultimate Edition*, 4 edition. Cambridge University Press.

Pingali Sailaja. 2012. Indian english: Features and sociolinguistic aspects. *Language and Linguistics Compass*, 6(6):359–370.

Edgar W. Schneider. 2007. *Postcolonial English: Varieties around the World*. Cambridge Approaches to Language Contact. Cambridge University Press.

Farzad Sharifian. 2020. *The Routledge Handbook of World Englishes*.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Benjamin Lee Whorf. 1956. Language, thought, and reality: selected writings of....(edited by john b. carroll.).

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

11 Appendix

Domain	**Alignments**
Quantity	0.6105
Kinship	0.4674
Time	0.4639
Possession	0.4264
Speech and language	0.4031
Miscellaneous function words	0.3931
Cognition	0.3787
Motion	0.352
Basic actions and technology	0.3298
Modern world	0.3239
Spatial relations	0.3235
Social and political relations	0.3213
Emotions and values	0.2737
Sense perception	0.2573
The body	0.2537
The house	0.2439
Clothing and grooming	0.2414
The physical world	0.24
Food and drink	0.1908
Animals	0.1681
Agriculture and vegetation	0.1133

Table 3: Mean Semantic Alignment for Indian and American English.

11.1 Code

`github.com/CodeOfCognition/Dialectal-Semantics`