# Global divergence and local convergence of utterance semantic representations in dialogue

**Yang Xu**
Department of Computer Science
San Diego State University
yang.xu@sdsu.edu

## Abstract

We use deep contextualized embedding models (BERT & ELMo) and shallow word embedding models (Fasttext & GloVe) to study the alignment between dialogue interlocutors at the semantic representation level, with the goal to examine the *interactive alignment model* (IAM) theory. We have observed both divergence and convergence patterns in dialogue: First, the semantic distance between two adjacent utterances increases with their relative positions within the dialogue, i.e., utterances at the later stage are more semantically apart than the earlier ones. Second, semantic distance also increases with the physical distance between utterances, i.e., utterances that are physically closer have more similar semantic meanings. We conclude that dialogue interlocutors demonstrate *global* divergence and *local* convergence patterns in semantic representation space. Our findings resolve the conflicts in previous studies, and challenge the claim from IAM that people gradually build alignment at higher representation levels in dialogue. The feasibility of using semantic representation techniques as psycholinguistic models of dialogue is discussed.

## 1 Introduction

How speakers in a dialogue adapt their language use from each other is a widely studied question. Many studies view dialogue as a dynamic process, whose temporal properties are closely examined. A commonly used paradigm is regressive analysis that examines how linguistic features at various representation levels change with time as the dialogue proceeds. One of the theoretical basis behind all these empirical exploration is the *interactive alignment model* (IAM) theory (Pickering and Garrod, 2004), which puts forward a hierarchical view of language adaptation: the alignment (i.e., re-use of same elements) at lower representation levels (e.g., phonetic, lexical) leads to the alignment at more abstract representation levels, such as *semantic* and *situation* representations. A straight-forward inference from this theory is that we should be able to empirically observe consistent evolving patterns (within dialogue) across representation levels.

However, studies from different approaches often have contradictory findings in the temporal patterns of dialogue. Many of the conflicts are about two concepts: *divergence* vs. *convergence*. For example, Healey et al. (2014) find that people diverge from their interlocutors in syntax use as dialogue develops. Xu and Reitter (2016a,b) find an opposite converging pattern in syntactic complexity and information density. Abney et al. (2014) find the convergence of timescale complexity in affiliative conversations, and divergence in argumentative ones. We deem that this inconsistency in existing findings is due to the lack of clear defining the *scale* under which alignment is examined. Multiple scales are needed in order to obtain a comprehensive understanding of alignment in dialogue.

Another gap in the existing studies is that most studies explore alignment at lower levels, such as phonetic, lexical, syntactic etc. Alignment at the higher representation levels in IAM is not well studied, such as semantic. With the development of representation techniques powered by deep neural networks, we are now able to obtain semantic representations for utterances that are more precise than before. It also means that the semantic relatedness, or distance, between dialogue participants can be easily measured now. Although whether similarity in semantic space is equivalent to the concept of "alignment" is still an open question, it is worth trying this new technique to extend the psychological theories of dialogue.

Therefore, in this study we use the distance of utterance semantic representations to measure the strength of semantic alignment, and focus on



*Proceedings of the Society for Computation in Linguistics (SCiL) 2021*, pages 116-124.
Held on-line February 14-19, 2021

the temporal property of alignment at two scales, global and local. First, we obtain the dense vector representation for each utterance using different neural semantic models. Then, we analyze how the distance between interlocutors in semantic space evolves within dialogue, i.e., whether and when convergence or divergence patterns can be observed. We include two types of dialogue corpora to ensure the generality of study.

## 2 Related Work

### 2.1 Divergence vs. convergence patterns in dialogue

The general principle put forward by IAM is that interlocutors become aligned as dialogue develops, which is a "convergence" perspective. Many existing studies have obtained similar or different findings. Healey et al. (2014)'s main finding is that people systematically diverge from each other in their use of syntactic structures. More specifically, they find that the syntactic similarity between immediate adjacent utterances (turns) is significantly lower than those that are farther apart. Their findings are essentially about the local temporal patterns of language. However, we argue that this result may also be caused by naturalistic "fluctuation" of *information density* in natural dialogue. Moreover, we argue that a comprehensive analysis based on sufficient empirical investigation on other linguistic representation levels (other than syntactic complexity) is needed, in order to draw general conclusions about the temporal characteristics of language in dialogue (Xu and Reitter, 2018).

Xu and Reitter (2016a,b) take a bigger scale by looking at the topic episodes in dialogue, and find convergence patterns in syntactic complexity and lexical information density. Abney et al. (2014) develops the concept of complexity matching, as an extension to behavioral alignment, and have observed convergence patterns within the affiliative dialogues. The analysis scale they take span across the whole dialogue. Thus, the scale used will affect the observed patterns and the conclusions about alignment.

To briefly summarize, Healey et al. (2014) uses a local scale on adjacent utterances within a fixed window, while Xu and Reitter (2016a)'s methods focus on a larger local scale spanning across more utterances, and Abney et al. (2014) takes a global scale that includes the whole body of dialogue. In order to have comprehensive understanding of linguistic alignment, however, multiple scales need be considered.

### 2.2 Semantic models of natural language

One recent big advancement in NLP is the development of deep contextualized representation models that can capture the rich semantic meanings of sentences and the constituent words. The term "contextualized" is reflected in their similar natures in modeling the context for each word "dynamically" using Transformers (Vaswani et al., 2017) and Long-Short-Term-Memory (LSTM) (Hochreiter and Schmidhuber, 1997). For example, ELMo (Embeddings from Language Models) (Peters et al., 2018) utilizes bidirectional two-layer LSTMs. BERT (Devlin et al., 2018) uses multiple layers of Transformers.

These new models take advantages of deep neural network architectures that can better capture complex characteristics of word use, such as syntax-semantics interaction, context-dependent meanings (polysemy) etc., which shallow models such as word2vec is unable to deal with. The essential difference from shallow models is that they assign each word a representation that is a function of the entire sentence (or sequence). But in order to have a comprehensive comparison, we still include two shallow word embedding models, Fasttext (Bojanowski et al., 2017) and GloVe(Pennington et al., 2014) in this study.

## 3 Method

### 3.1 Corpus data

We examine two types of dialogue in this study, *spontaneous* and *task-oriented*.

**Type 1: Spontaneous**. The Switchboard corpus (Godfrey et al., 1992) and the British National Corpus (BNC) (BNC, 2007) are used in this study. Switchboard contains 1126 dialogues by telephone between two native North-American English speakers in each dialogue. We use only a subset of BNC (spoken part) that contain spoken conversations with exactly two participants, so that the dialogue structures are consistent with Switchboard. BNC contains both written and spoken texts, and the spoken texts further consists of two parts: the demographically sampled part (BNC-DEM), which contains impromptu speech in informal settings, and the context-governed part (BNC-CG), which is sampled from more formal settings (Tottie, 2011). To be consistent with the Switchboard corpus and

to simplify our experiment as well, we select parts of BNC-DEM and BNC-CG that only have two speakers within each conversation, which contains 1346 dialogues in total. For convenience, in later part of this paper, we simply use BNC to refer to this sampled part of BNC-DEM.

**Type 2: Task-oriented**. Two corpora are examined in this study: the HCRC Map Task Corpus[1] (Anderson et al., 1991) (Maptask) and a smaller corpus in Danish from a joint decision-making study (Fusaroli et al., 2012) (henceforth DJD). Maptask contains a set of 128 dialogues between two subjects, who accomplished a cooperative task together. DJD contains a set of 32 dialogues from native speakers of Danish collected by (Fusaroli et al., 2012). An overview of the four corpora are shown in Table 1.

### 3.2 Semantic models

We compare two methods of obtaining the semantic representations of utterances: aggregated word embeddings, and contextualized sentence embeddings.

**Method 1: Aggregated word embeddings**.
This method computes the summation of the embedding vectors of all the tokens (words) within an input sequence (sentence) normalized by the sequence length. It has been widely used to approximate the semantic meanings of sequences (sentences). It is often used as a baseline in comparison with more sophisticated sequence models, for tasks such as text classification and information retrieval etc (Wieting et al., 2015). Simple this method may seem, we deem it a reasonable model to start with. In particular, we choose two types of pretrained word embeddings, GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), because of their availability for multiple languages. A summary of the parameters of the pretrained models used is shown in Table 2.

**Method 2: Contextualized sentence embeddings**.
This method uses more sophisticated models to obtain richer semantic representations of input sequences. Two models are used, BERT and ELMo. In particular, we use the pretrained BERT model[2] and ELMo embeddings[3]. We choose ELMo as opposed to other models (e.g., GPT etc.) based on the following consideration: first, it is a typical non-transformer-based model, which supports fast inference; second, the availability of pre-trained model in Danish.

**Semantic distance**
We measure the semantic distance between two utterances using the cosine distance between their representation vectors. An utterance $u$ can consist of multiple sentences, $u = \{s_1, s_2, \dots s_N\}$ ($N \geq 1$). Each sentence $s_i$ has a semantic representation $r_i$ either from Method 1 or 2, and we use the simple average of them as the semantic representation of the utterance, $R_u = \frac{1}{N}\sum_{i=1}^{N} r_i$.

Given two target utterances $u_A$ and $u_B$, we first obtain their representation vectors, $R_{u_A}$ and $R_{u_B}$, and then compute their cosine distance,

$$\textit{semDist}(R_{u_A}, R_{u_B}) = 1 - \frac{R_{u_A} \cdot R_{u_B}}{\|R_{u_A}\|\|R_{u_B}\|}$$

The value of *semDist* is within the range of $[0, 2]$. A larger value means that the two representation vectors are farther away in the semantic representation space, i.e., the two utterances are more semantically apart. A smaller value indicates the opposite, i.e., they are more semantically similar.

Therefore, there are in total $4 \times 4 = 16$ model-corpus combinations: {BERT, ELMo, Fasttext,GloVe} $\times$ {Maptask, DJD,Switchboard, BNC}. The only one exception is in DJD corpus: we did not find any pretrained GloVe model in Danish, so we used a Word2vec (Mikolov et al., 2013) model of the same size instead.

### 3.3 Defining key variables

We give operational definitions to the concepts of "adjacent utterances" and "timestamps", which are necessary in order to analyze the temporal patterns in dialogue. Given a pair of utterances $p = \langle u_A, u_B \rangle$, we consider two key quantities: first, the "physical" distance between them (*phyDist*) in terms of utterance count; second, the relative position of $p$ within the dialogue, (*relPos*). *phyDist* can be used to model the strength of priming-induced alignment between interlocutors, while *relPos* models the "timestamps" of utterance pairs, which characterizes the overall evolving pattern of alignment.

First, we assign an integer index $idx$ ($idx = 1, 2, \dots$) to each utterance in dialogue. Utterance is identified based on speech turns, i.e., whenever

[1] http://groups.inf.ed.ac.uk/maptask/
[2] With this opensource tool: https://github.com/huggingface/transformers
[3] With this opensource tool: https://github.com/HIT-SCIR/ELMoForManyLangs

| Corpus | Dialogue count | Token count | Avg dialogue length (SD) | Avg utterance length (SD) |
|---|---|---|---|---|
| Switchboard | 1126 | 1.61 M | 106.4 (49.6) | 13.4 (17.2) |
| BNC | 1346 | 1.27 M | 58.9 (112.1) | 16.0 (53.4) |
| Maptask | 128 | 194.2 K | 147.4 (72.6) | 10.3 (9.4) |
| DJD | 32 | 65.9 K | 269.5 (113.2) | 7.6 (5.9) |

Table 1: Overview of the four corpora used. Dialogue length (3rd column) is measured with the number of utterances in a dialogue; utterance length (4th column) is measured with the number of words in an utterance.

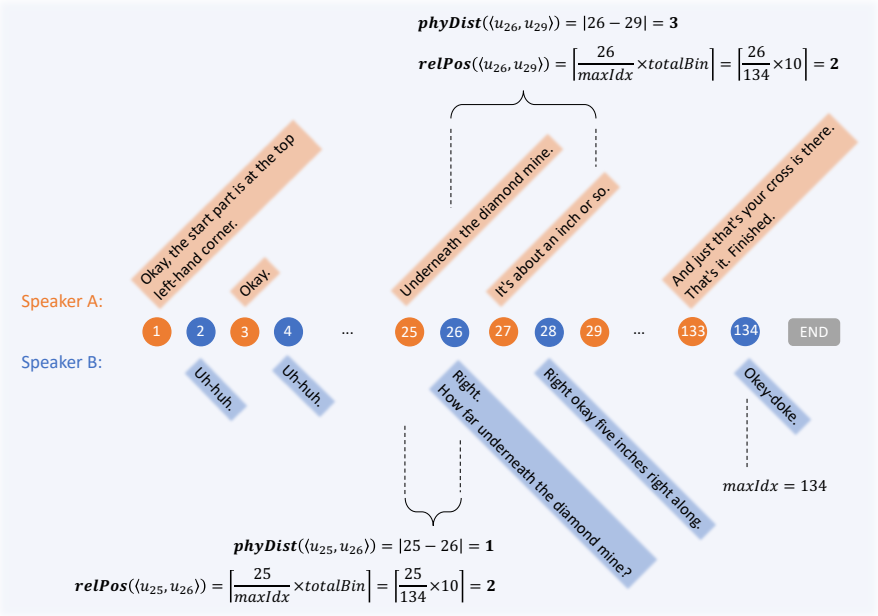


Figure 1: Demonstration of how the two keys variables, *relPos* and *phyDist*, are defined. In the example, the utterance pair $\langle u_{25}, u_{26} \rangle$ and $\langle u_{26}, u_{29} \rangle$ have the same *relPos* value, but are of different *phyDist* values.

| Model | Params & training data |
|---|---|
| BERT | 12-layer, 768-hidden, 12-heads trained on Wikipedia data |
| ELMo | 2-layer, 1024-hidden trained on 20M tokens data |
| GloVe | 300-dim trained on 6B tokens data |
| FastText | 300-dim trained on 42B tokens data |

Table 2: Parameters of the pretrained semantic models.

a turn-taking occurs, the upcoming consecutive sequence of tokens is considered as a new "utterance" (regardless of its length). Thus, short back-channel utterances are also included in our initial analysis. Then, the *phyDist* of utterance pair $\langle u_A, u_B \rangle$ is defined as the subtraction between the indices of $u_A$ and $u_B$,

$$phyDist(\langle u_A, u_B \rangle) = |idx(u_A) - idx(u_B)|$$

Because all the corpora data we analyze consist exactly of two participants, who take turns to talk, subtracting the two indices from any utterance pairs from two different speakers must result in an odd number: $1, 3, 5, \ldots$ (An even subtraction means that the two utterances are of the same speaker ship, i.e., self-alignment, which is not the focus of this study). *phyDist* can be arbitrarily big, as long as it is smaller than the dialogue length. But according to the theory that alignment is due to

short-term priming effect, we only need to consider utterance pairs that are within a reasonable distance. Thus, we limit the values of *phyDist* to the set of $\{1, 3, 5, 7, 9\}$.

To define the *relPos* of $\langle u_A, u_B \rangle$, we first divide all utterances in a dialogue into ten (10) bins of equal window size, and each utterance pair falls into one of the ten bins. Then the *relPos* of a pair is defined as the bin number to which it belongs, as follows,

$$\textit{relPos}(\langle u_A, u_B \rangle) = \left\lceil \frac{idx(u_A)}{\textit{maxIdx}} \times \textit{totalBins} \right\rceil$$

where *maxIdx* is the $idx$ of the last utterance in dialogue (i.e., total # of utterances), and *totalBins* is 10 in this case. Note that $idx$ encodes the position information of an utterance, and *relPos* is a normalized version of it (normalized by dialogue length) for the purpose nicer visualization. In the later part of the paper, $idx$ will still be useful in fitting statistical models. Because we set $totalBins = 10$, the range of *relPos* is also $\{1, 2, 3, \ldots, 10\}$. A demonstration of how *phyDist* and *relPos* are defined is shown in Figure 1.

## 4 Results

### 4.1 Semantic distance increases with relative position

We first study the alignment of semantic representations at the global scale. We plot the semantic distance (*semDist*) between utterance pairs against their relative positions (*relPos*) within dialogue (Figure 2). It can be seen that for 14 out of the 16 model-corpus combinations, *semDist* increases with *relPos*, with the only two exceptions, $\{\text{Fasttext}, \text{GloVe}\} \times \text{BNC}$. To examine the statistical reliability of the observed increasing trend in the plots, we fit a linear mixed-effect model for each one of the 16 model $\times$ corpus combinations. The model uses *semDist* as the dependent variable, utterance $idx$ as the fixed effect predictor, and with random intercepts fitted per dialogue. As explained in Section 3.3, we use $idx$ instead of *relPos* as the predictor, because the former encodes more fine-grained position information, while the latter is used mainly for visualization purpose. The coefficients and significance levels of the 16 individual models (for each model-corpus combination) are shown in Table 3. The statistical significance of the observed increasing trend can be confirmed.

### 4.2 Semantic distance increases with physical distance

Next, we study the alignment at local scale, by plotting *semDist* against (*phyDist*) (Figure 3). An obvious increasing trend of *semDist* with *phyDist* can be observed in 3 out of the 4 corpora: Maptask (except ELMo), DJD, and BNC. A decreasing trend is found in Switchboard. Similarly, we fit linear mixed-effect models to confirm the statistical significance (See Table 4).

### 4.3 Short utterances removed

The major discrepancy in the above shown results is the distinctive decreasing pattern of Switchboard corpus across all models in the *semDist* $\sim$ *phyDist* relationship. This is somewhat surprising because it is anti-intuitive to see utterances farther apart have closer meanings (see the discussion in Section 5.2). Here, we examine whether removing those short utterances from data will produce more consistent results.

One character of dialogue transcript data is the frequent usage of extremely short back channels utterances and disfluencies, such as "uh", "um", "hmm" etc. The meanings of these short utterances are vague and unrepresented in the pretrained semantic models, because the training data of these models are usually in well-formed written language. We find that Switchboard has the most imbalanced distribution, with the highest proportion of short utterances. It could be a cause to the discrepancy in the previous result.

To examine the assumption, we conduct an experiment by excluding the utterances that are below certain thresholds, and re-examine the *semDist* $\sim$ *phyDist* relationship in the remaining data. We find that after removing the utterances that contain fewer than 2 tokens (in Switchboard), the original decreasing trend of *semDist* against *phyDist* is reversed (now becomes *increasing*) for Fasttext and GloVe (See Figure 4), but not so for BERT or ELMo. This tells that the behaviors of semantic models can indeed be affected by the nature of data. The negative correlation between *semDist* and *phyDist* can be observed in Switchboard, as long as short utterances are removed. But we wonder it does not work for BERT or ELMo, and we leave it to future investigation.

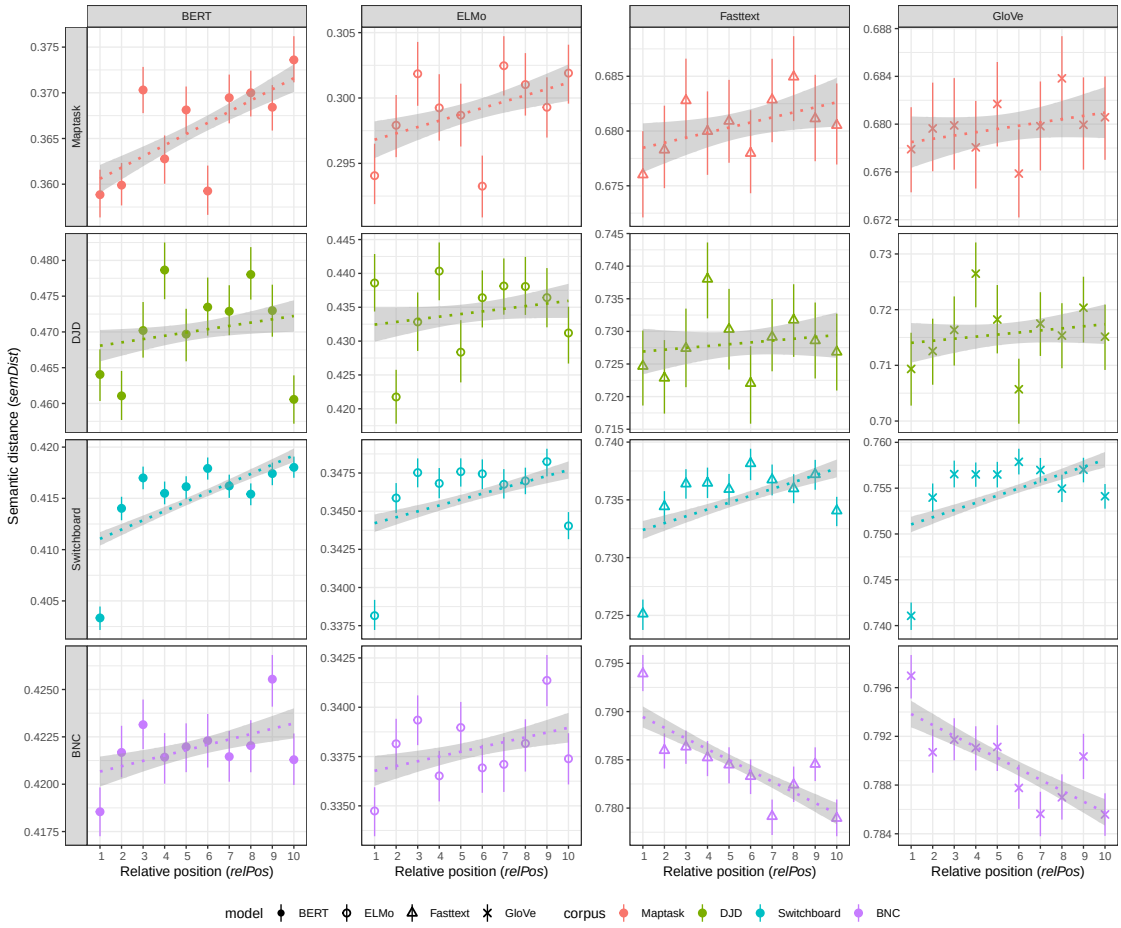


Figure 2: Semantic distances between utterance pairs (*semDist*, $y$-axis) against their relative positions (*relPos*, $x$-axis) for 16 model $\times$ corpus combinations. Results from the same model is in one column, and those from the same corpus is in one row. Dotted lines indicate fitted simple linear models. Lines across points and shaded areas indicate 95% confidence intervals.

## 5 Discussion

### 5.1 Global semantic divergence

The results in Section 4.1 is inconsistent with the claim from IAM that interlocutors become aligned semantically as dialogue develops, nor with some existing empirical findings (Abney et al., 2014). The "divergence" we observed here is different from the one reported by Healey et al. (2014), because they use a local scale of fixed window, which really means local divergence.

Here, we try to explain the seemingly counter-intuitive global divergence. In the case of spontaneous (non-task-oriented) dialogues (e.g., Switchboard and BNC), what people talk about typically will disperse, which naturally leads to the divergence of utterance meanings. Even though in the collection of some corpora, participants are asked to speak with respect to certain topics (e.g., Switchboard), a clear trend of topic shifting can still be detected computationally (Xu and Reitter, 2016b). In the case of task-oriented dialogue, such as Maptask, interlocutors tend to first establish a common "language" (words, terms, etc.) to describe the situation, i.e., *common ground* (Clark, 1996), which will be the basis of later communication. Our findings suggest that although common ground is necessary, this grounding or alignment process may not be directly reflected on to the semantic representation of utterances.

Our findings inevitably raise some issues in the expressions of IAM theory: it seems semantic convergence (measured by semantic distance) is not observed among most dialogue corpora, so what is it that has become aligned between interlocutors? Is it something more abstract that is difficult to

| Corpus \ Model | BERT | ELMo | Fasttext | GloVe |
|---|---|---|---|---|
| Maptask | $5.4 \times 10^{-5***}$ | $2.1 \times 10^{-5**}$ | $3.9 \times 10^{-5***}$ | $3.0 \times 10^{-5**}$ |
| DJD | $1.9 \times 10^{-5**}$ | $1.1 \times 10^{-5}$ | $1.6 \times 10^{-5}$ | $2.3 \times 10^{-5*}$ |
| Switchboard | $4.6 \times 10^{-5***}$ | $1.6 \times 10^{-5***}$ | $1.7 \times 10^{-5***}$ | $2.8 \times 10^{-5***}$ |
| BNC | $1.1 \times 10^{-5***}$ | $1.1 \times 10^{-5***}$ | $-2.1 \times 10^{-5***}$ | $-1.8 \times 10^{-5***}$ |

Table 3: $\beta$ coefficients and the significance levels of utterance $idx$ as the predictor in linear models: *semDist* $\sim$ *idx*. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

| Corpus \ Model | BERT | ELMo | Fasttext | GloVe |
|---|---|---|---|---|
| Maptask | $3.6 \times 10^{-4**}$ | $-5.4 \times 10^{-4***}$ | $8.0 \times 10^{-4***}$ | $1.1 \times 10^{-4***}$ |
| DJD | $1.0 \times 10^{-3***}$ | $1.5 \times 10^{-3***}$ | $2.7 \times 10^{-3***}$ | $2.7 \times 10^{-3***}$ |
| Switchboard | $-4.4 \times 10^{-3***}$ | $-3.1 \times 10^{-3***}$ | $-2.1 \times 10^{-3***}$ | $-3.2 \times 10^{-3***}$ |
| BNC | $8.1 \times 10^{-5}$ | $1.1 \times 10^{-4}$ | $2.0 \times 10^{-3***}$ | $1.8 \times 10^{-3***}$ |

Table 4: $\beta$ coefficients and the significance levels of *phyDist* as the predictor in linear models: *semDist* $\sim$ *phyDist*. $^{***}p < .001$, $^{**}p < .01$.

capture by vector-based models? These are also open-ended questions to future investigations.

### 5.2 Local semantic convergence

The local convergence pattern is in line with most computational evidence about linguistic alignment. Reitter et al. (2006); Reitter and Moore (2014) find that the priming effect of words and syntactic structures decay over the distance between utterances, and they use the decay rate to model the strength of alignment. Similarly, our findings show that the semantic similarity (opposite of distance) between utterances also decays as they get apart. We conjecture that this observation can be partially (a small part) due to decay effect of lexical alignment, but it is more of a direct reflection of the natural process in dialogue. Utterances in dialogue are mostly the replies to adjacent ones, and it makes sense that nearby utterances are more semantically related. It is also a specification of the more general *temporal clustering*, or *burstiness* property (Goh and Barabási, 2008; Jo et al., 2012) that pervasively exists in natural time series.

Our results contradicts those from Healey et al. (2014)'s work. As mentioned before, the syntactic divergence they reported is in fact a local divergence. It at least indicates that more overlaps of syntax in utterance does not necessarily co-occur with high semantic similarity.

## 6 Conclusions and Future Work

The main contribution of this study is to work towards a more complete theory of the linguistic alignment between dialogue interlocutors, by examining the alignment at semantic level with the most recent representation techniques. Our findings suggest that interlocutors in dialogue consistently diverge from each other in semantic space at the *global* scale; and conversely, they converge at the *local* scale. This mixture of scales we use is a novel perspective of exploring inter-speaker alignment.

Our findings help clarify the "convergence vs. divergence" conflicts in previous studies, and raises some limits in IAM theory. In particular, we challenge the belief that dialogues converge towards some aligned semantic space. Rather, the semantic meanings of utterances diverge globally in a dialogue, while the local convergence pattern is maintained.

For future work, we are considering fine tuning the semantic models on dialogue data, instead of just using pretrained models, with the hope to get more accurate semantic representation of utterances.

## Acknowledgments

We thank all the reviewers for providing the valuable comments and thoughts. We also thank the conference organizers for their effort in keeping

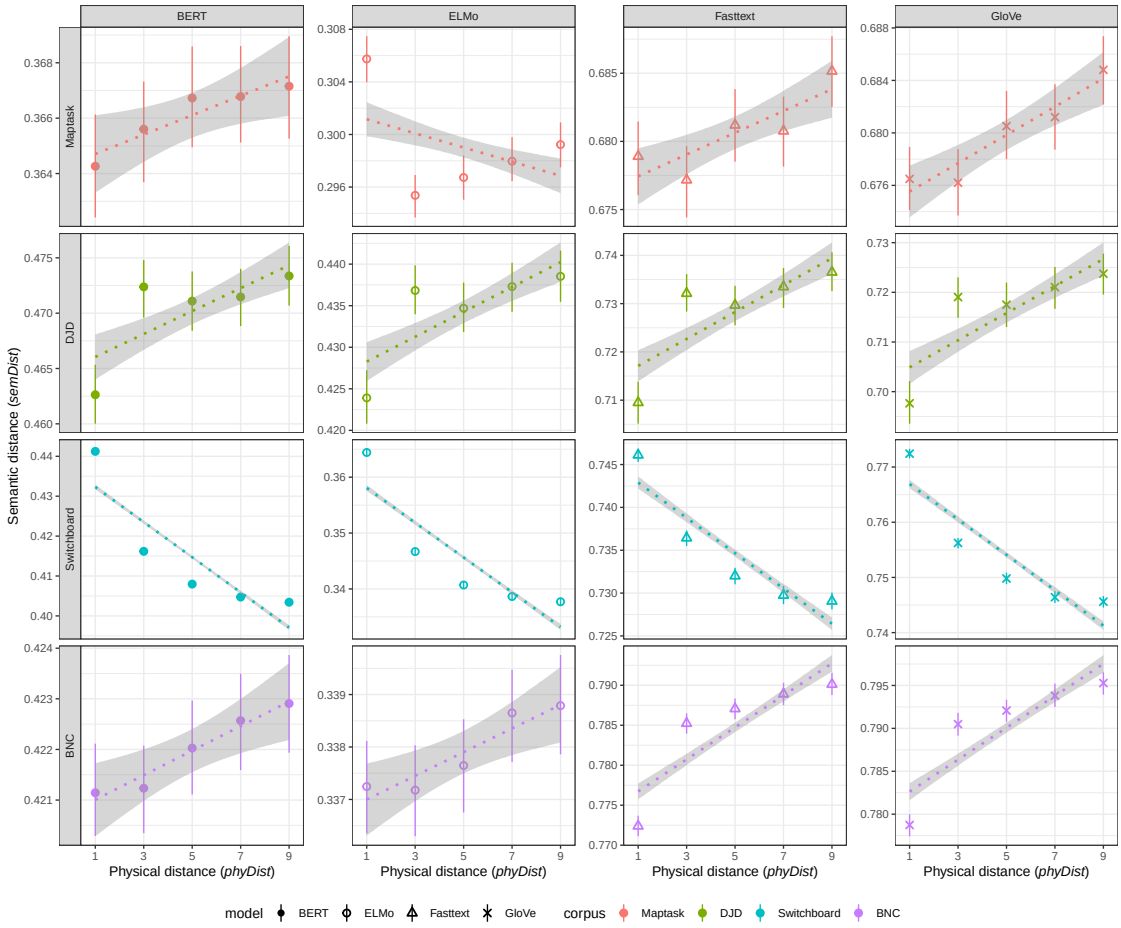


Figure 3: Semantic distance between utterance pairs (*semDist*, $y$-axis) against the physical distance (*phyDist*, $x$-axis) for 16 model $\times$ corpus combinations. Results from the same model is in one column, and those from the same corpus is in one row. Dotted lines indicate fitted simple linear models. Lines across points and shaded areas indicate 95% confidence intervals.

everything running during this hard time that we are all going through. Thank you and solute!

## References


Drew H Abney, Alexandra Paxton, Rick Dale, and Christopher T Kello. 2014. Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, 143(6):2304.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

BNC. 2007. The British National Corpus, version 3 (BNC XML Edition).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Herbert H Clark. 1996. *Using Language*. Cambridge university press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to terms quantifying the benefits of linguistic coordination. *Psychological Science*, 23(8):931–939.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

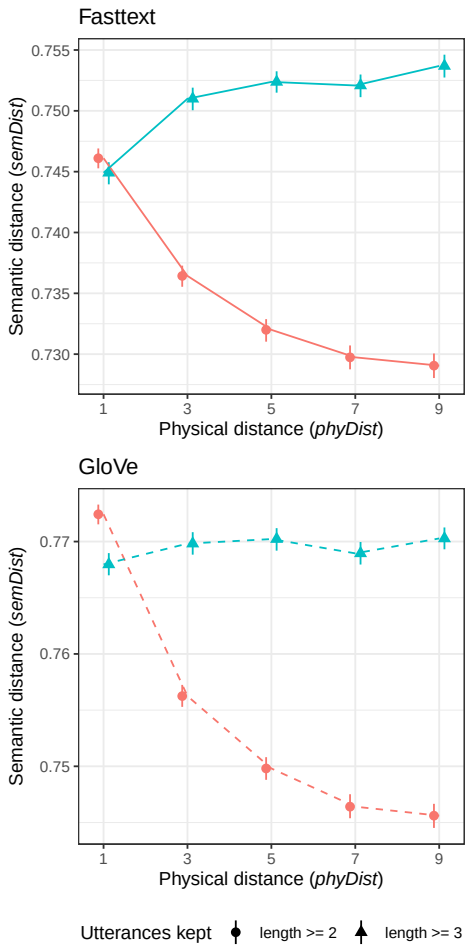


Figure 4: The *semDist* $sim$ *phyDist* correlation changes direction when short utterances (length $\leq$ 2) are removed in Switchboard. This flip of trends only holds to Fasttext and GloVe models.


K-I Goh and A-L Barabási. 2008. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS One*, 9(6).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. 2012. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York City, USA. Association for Computational Linguistics.

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Gunnel Tottie. 2011. Uh and um as sociolinguistic markers in british english. *International Journal of Corpus Linguistics*, 16(2):173–197.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Yang Xu and David Reitter. 2016a. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448.

Yang Xu and David Reitter. 2016b. Entropy converges between dialogue participants: explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–546.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.