An Evaluation and Comparison of Linguistic Alignment Measures

Yang Xu and **David Reitter**
College of Information Science and Technologies
The Pennsylvania State University
University Park, PA 16802, USA
yang.xu@psu.edu, reitter@psu.edu

Abstract

Linguistic alignment has emerged as an important property of conversational language and a driver of mutual understanding in dialogue. While various computational measures of linguistic alignment in corpus and experimental data have been devised, a systematic evaluation of them is missing. In this study, we first evaluate the sensitivity and distributional properties of three measures, indiscriminate local linguistic alignment (LLA), Spearman's correlation coefficient (SCC), and repetition decay (RepDecay). Then we apply them in a study of interactive alignment and individual differences to see how well they conform to the Interactive Alignment Model (IAM), and how well they can reveal the individual differences in alignment propensity. Our results suggest that LLA has the overall best performance.

1 Introduction

The alignment of language between dialogue partners has garnered much interest in the computational linguistics community. Alignment not only provides insight into the mechanisms of dialogue, but also has the potential to improve both human-computer dialogue systems and the analysis tool-chain. In this context, alignment refers to the convergence of linguistic choices among interlocutors. This may happen at different representational levels, such as the phonological, lexical and syntactic (Garrod and Anderson, 1987). Alignment, also known as entrainment or accommodation, has become recognized as a key feature of linguistic communication.

Several theoretical accounts exist that address the nature and implications of linguistic alignment. In psycholinguistics, the Interactive Alignment Model (IAM) assumes that interlocutors align their linguistic representations (Pickering and Garrod, 2004), from lower ones (lexical, syntactic) to higher ones (e.g., semantics), leading to shared situation models. Sociolinguistic studies point out that interactants converge in their communication styles to signal social affinity and diverge to emphasize social distance (Danescu-Niculescu-Mizil and Lee, 2011; Giles, 2008). Furthermore, evidence has been found showing that certain individuals tend to have higher propensity of alignment than others (Gnisci, 2005; E. Jones et al., 1999; S. Jones et al., 2014; Willemyns et al., 1997).

Several computational measures have been developed to help validating these theoretical accounts. Some of them use the probability of co-occurrence of words (or other linguistic elements) to describe the language alignment (Church, 2000; Dubey, Sturt, and Keller, 2005; Reitter, Keller, and Moore, 2006), while some others take inspiration from documents similarity measures (Huffaker et al., 2006; S. Jones et al., 2014; Wang, Reitter, and Yen, 2014).

However, little research is available that evaluates the properties of these linguistic alignment measures. How sensitive are these measures? What kind of distributions do they have? Can they consistently describe the alignment at multiple linguistic levels (e.g., lexical and syntactic)? Can they describe the individual differences in propensity of alignment? Essentially, are they good/reliable measures? These questions are not answered (or fully answered) yet.

To answer these questions in this study, we first conduct an evaluation of the intrinsic properties of three well defined and commonly used measures, indiscriminate local linguistic alignment (LLA) (Fusaroli et al., 2012; Wang, Reitter, and Yen, 2014), Spearman's correlation coefficient (SCC) (Huffaker et al., 2006; Kilgarriff, 2001), and repetition decay (RepDecay) (Reitter, Keller, and Moore, 2006), in which two basic properties are investigated, normality of distribution and sensitivity. Then we apply these measures to a study about the IAM and individual differences in alignment propensity as an extrinsic evaluation. We examine how well they follow the basic assumption of IAM, i.e., showing correlations between alignment at lexical and syntactic levels, and how well they can reveal the individual differences in alignment propensity.

Our study aims to provide potential guidance to future studies of linguistic alignment in terms of which computational measures to use. Basically, we favor a measure that has good normality in its distribution, that has higher sensitivity, and that conforms with the IAM theory and the existing findings about individual differences in alignment propensity.

2 Related Work

We will first briefly review the existing computational measures of linguistic alignment. Then we give a short reivew of IAM and the work on individuals' propensity of alignment.

2.1 Existing measures and their limitations

We categorize the existing computational measures into three basic types based on the different methods they use. Though different methods are used, all the three types of measures are conducted upon a similar structure: ($prime$, $target$) pairs, in which $prime$ and $target$ are pieces of text.

Probabilistic measures

Probabilistic measures work on multiple ($prime$, $target$) pairs, and compute the probability of a single word or syntactic rule appearing in $target$ after its appearance in $prime$, by counting the frequency of their co-occurrence. For example, Church (2000) used the first half of documents as $prime$ and the second half as $target$ to measure the lexical adaptation in text. Dubey, Sturt, and Keller (2005) used similar measures to investigate the parallelism effect of syntactic structures in coordinate constructs in corpora. Gries (2005) was among the first to use logistic regression to estimate linear models of syntactic priming.

The limitation of the frequency-based measure is that it needs a relatively large amount of text to conduct the computation, because it uses the observed frequency of words (or syntactic rules), to estimate the probability of co-occurrence.

Document similarity measures

Several measures originate from information retrieval (IR). They have seen little use by corpus-based priming and alignment researchers, although they could conceivably be adopted for our purposes. Huffaker et al. (2006) compared the performance of three computational measures of document similarity in measuring the language convergence in an online community over time. The measures they examined are: Spearman's correlation coefficient (SCC), which measures document similarity based on word frequency and co-occurrence, Zipping, a data compression algorithm that has been used in document comparison, and Latent Semantic Analysis (LSA), a technology for measuring semantic similarity between documents.

Fusaroli et al. (2012) proposed a measure based on probabilities that falls in this category as well: the concept of indiscrimiate local linguistic alignment (LLA). Based on this work, Wang, Reitter, and Yen (2014) implemented LLA at lexical level (LILLA) and syntactic level (SILLA). They essentially measure the number of words (or syntactic rules) that appear in both $prime$ and $target$, normalized by the size of the two text sets.

Repetition decay

Repetition effects have been observed to be short-lived in experiments (e.g., Branigan, Pickering, and Cleland, 1999). Reitter, Keller, and Moore (2006) proposed to use the decay rate of repetition probabilities of syntactic rules to measure the strength of syntactic alignment, and to apply it to all syntactic rules in an observational study.

In their work, Reitter, Keller, and Moore (2006) built a generalized linear model, using the repetition of the syntactic rules as the dependent variable and the distance between $prime$ and $target$ as the predictor. They observed that repetition rate of syntac-

tic rules decays as the distance increases, and used the regression coefficient of the predictor to estimate the strength of syntactic alignment.

Repetition decay gives a strict mathematical account to the alignment phenomena from the probabilistic point of view, and distinguishes the alignment caused by priming from other random repetitions of linguistic elements. One limitation of the repetition decay measure is that it cannot quantify the alignment between a single pair of texts (in fact, it assumes that the simple repetition between two text sets tells us nothing about the overall alignment level). Another limitation is that the fitting a generalized linear model is not as computationally efficient as other measures.

2.2 Interactive alignment model

Pickering and Garrod (2004) proposed the Interactive Alignment Model (IAM) to account for the mechanism that underlie language processing in dialogue. The central assumption of IAM is that, in a dialogue, the linguistic representations employed by the interlocutors become aligned at many levels, and the aligned representations at one level lead to aligned representations at other levels (Pickering and Garrod, 2004). The correlation between different linguistic levels has been shown by corpora-based studies (Wang, Reitter, and Yen, 2014).

2.3 Propensity of alignment

One area that has long been overlooked is the individual speaker's inherent propensity of alignment, i.e., whether some individuals inherently have a stronger tendency to align to their interlocutors than others. Previous studies have shown that individuals in lower social power status tend to converge their language style to those in higher social power status during conversations, e.g., interviewees converging towards their interviewers during employment interviews (Willemyns et al., 1997), students adapting their language to teachers (E. Jones et al., 1999), and witnesses accommodating their linguistic style to that of the lawyers and the judges (Gnisci, 2005). More recently, S. Jones et al. (2014) proposed *Zelig Quotient*, a measure that characterizes an individual's inherent tendency to accommodate to the linguistic style of others, defined by the movement in a high-dimensional linguistic style space. These studies provide evidence that different individuals have different levels of alignment propensity, and this difference can be quantified by computational measures.

However, the main limitation of existing studies is that the individuals' propensity of alignment is only characterized using a proportion of lexical elements. For example, Zelig Quotient only uses functional words (S. Jones et al., 2014). Thus they do not characterize the propensity of alignment at the full range of lexical and syntactic levels.

3 Evaluation Criteria

In this study, we first evaluate two intrinsic properties of the computational measures, and then evaluate their performance in two extrinsic investigations related with IAM and individuals' propensity of alignment.

3.1 Intrinsic evaluation

The two intrinsic properties that we find desirable are: normality of distribution and sensitivity. We expect a good measure to have a normal (or nearly normal) distribution over the whole population, because normal distribution is the most common distribution in nature, and it is desirable from a statistical point of view to have a normal distribution. The sensitivity criterion is straight-forward: we expect a good measure to have satisfactory "resolution", i.e., the capability of detecting relatively small amount of linguistic alignment.

3.2 Extrinsic evaluation

According to the IAM, linguistic alignment between interlocutors occurs at many levels, and aligned representations at one level leads to aligned representations at other levels. For instance, syntactic alignment is enhanced when there are more shared lexical items (Pickering and Garrod, 2004). Thus, it is reasonable to expect that a good measure can capture this effect, demonstrating that higher lexical alignment should co-occur with higher syntactic alignment.

Secondly, due to the empirical evidence that demonstrates the individual's inherent propensity of alignment (Gnisci, 2005; E. Jones et al., 1999; S. Jones et al., 2014; Willemyns et al., 1997), it is reasonable to expect that a good measure of linguis-

tic alignment should be able to characterize an individual's propensity of alignment. If we view the propensity of alignment as a relatively stable individual characteristic that is associated with other social and psychological factors, a good measure should be able to show more variation when measuring text produced by different individuals, and show less variation when measuring text produced by the same individual.

In sum, for the evaluation of the measures' intrinsic properties, we have two criteria: the *normality* of distribution and the *sensitivity*. For the extrinsic evaluation, we examine the performance of measures in three aspects: *consistency*, the measures at lexical level should be correlated with the measures at syntactic level. *Between-individual difference*, whether the measure can reveal significant differences in alignment propensity among different individuals. *Within-individual stability*, whether the alignment measures from the same individual have relatively small variance.

4 Methods

4.1 Processing of corpora

Four corpora are used in this study, including the text data from two online forums, the Cancer Survivors' Network (CSN) [1] and a massive open online course on visual art Art taught on Coursera by Penn State (MOOC), and two published corpora, the Switchboard Corpus (SWBD) (Marcus et al., 1994) and the spoken part of British National Corpus (BNC, 2007).[2]

The threads in CSN and MOOC have similar structures. They consist of an original post followed by reply posts ordered by time. We use a sequence of posts to represent a thread of length n, $[P_0, P_1, P_2, ..., P_n]$, in which P_0 represents the original post started by a forum user, and $P_i (i = 1, ..., n)$ represent the reply posts from other users or the original poster. There is a "reply" relationship between the posts in a thread, indicating that one post is a response to another. For example, if post j (by user B) is a "reply" to post i (by user A), then it means that post j is the direct response from user B to user A in terms of the content of post i. We construct the ($prime$, $target$) pairs for the linguistic alignment measures based on the "reply" relationship between the posts, i.e., using the original post as $prime$, and the corresponding reply post as $target$. Those pairs of posts whose authors are the same user ("self-reply") are excluded.

Switchboard has only two interlocutors in each conversation, whose utterances are ordered by turn. In BNC, one conversation might contain more than two interlocutors, which results in the relative loose structure of the conversation. The ways we construct ($prime$, $target$) pairs for the two corpora are similar: selecting one utterance as $prime$, and all the following utterances (within the distance of 10 utterances) that are from the other speaker are selected as $target$ respectively. We restrict the distance to 10 to avoid overtly long conversations. In total, we use all the 80,000 utterances in SWBD and randomly sample 95,441 conversations from BNC.

4.2 LLA

We use the methods implemented by Wang, Reitter, and Yen (2014) to compute the indiscriminate local linguistic alignment (LLA). The lexical and syntactic versions of LLA are implemented and abbreviated as LILLA and SILLA respectively. LILLA and SILLA are the normalized measures of the number of words (or syntactic rules) that occur in both the prime text and the target text:

$$\text{LLA}(P, T) = \frac{\sum_{w_i \in P} \delta(w_i, T)}{length(P) * length(T)} \quad (1)$$

$$\delta(w_i, P) = \begin{cases} 1, \text{if } w_i \in P \\ 0, \text{otherwise} \end{cases} \quad (2)$$

For the computation of LILLA, $|P|$ and $|T|$ are the numbers of words in $prime$ and $target$, and w_i is the individual word in $prime$ (or $target$). For the computation of SILLA, we first use the Stanford Parser (De Marneffe, MacCartney, Manning, et al., 2006) to parse each sentence in $prime$ and $target$ to get their full syntax trees, and then collect all the sub-trees from each sentence. For example, if the first sentence in $prime$ is "I am a teacher.", then the

[1] http://csn.cancer.org

[2] CSN has more than 48,000 threads collected in over 10 years. Switchboard contains more than 80,0000 transcribed utterances annotated with phrase structure trees (Marcus et al., 1994). We use 200 randomly sampled, spontaneous, multi-party conversations from BNC.

parser generates the full syntax tree: (S (NP (PRP I)) (VP (VBP am) (NP (DT a) (NN teacher)))). The sub-trees extracted are: "S $\rightarrow$ NP + VP", "NP $\rightarrow$ PRP", "VP $\rightarrow$ VBP + NP", "NP $\rightarrow$ DT + NN". Then we use the collection of all the sub syntax trees from $prime$ and $target$ as the $|P|$ and $|T|$ in Equation 1, and let w_i refer to the individual syntactic rules.

Differing from Wang, Reitter, and Yen (2014)'s work, we use the natural logarithm of LILLA and SILLA instead, i.e., log-LILLA and log-SILLA, as a simple way to achieve normality of errors.

4.3 Spearman's correlation coefficient

Spearman's correlation coefficient (SCC) originate from the Spearman rank correlation that has been widely used in statistics. It is essentially a non-parametric version of Pearson's correlation coefficient (Myers, Well, and Lorch, 2010). SCC was first proposed by Kilgarriff (2001) to measure the similarity between text and further evaluated by Huffaker et al. (2006). Huffaker et al. (2006) implemented SCC as the following: given a document pair ($prime$, $target$), for each document, rank the n common words in $prime$ and $target$ by frequency. For each word, let d be the difference of ranks in two documents. SCC is defined as the normalized sum of squared differences:

$$\text{SCC} = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{3}$$

SCC was originally implemented only for measuring the similarity at lexical level. In this study, we also implement the syntactic version of SCC by applying equation (3) to syntactic rules instead of words, i.e., first parse the $prime$ and $target$ into syntactic rules and get a list of common rules between the two sets, and then compute d in a similar way. In this study, we name the syntactic version of SCC as SCC_{syn}, and the original lexical version as SCC_{lex}.

4.4 Repetition decay

We compute the repetition decay (RepDecay) measure based on the procedure proposed by Reitter, Keller, and Moore (2006). We go through the sequence of ($prime$, $target$) pairs constructed from the corpora with a window of fixed width, e.g., 10 posts/utterances, and look at every element (a word or a syntactic rule) that is in $target$. If one element is also in $prime$, we record this in the variable Rep as 1, and otherwise, we record Rep as 0. Meanwhile, each Rep is associated with another variable $Dist$, which records the distance (from 1 to 10) between $prime$ and $target$. Finally, we build a generalized linear regression model using Rep as the response variable and $ln(Dist)$ as the predictor. We use the regression coefficient β associated with $ln(Dist)$ to represent the strength of linguistic alignment. Theoretically, β is always negative, and the smaller β indicates stronger alignment.

The computation of RepDecay relies on the precise definition of distance between $prime$ and $target$, because its basic assumption is that the priming effect from $prime$ to $target$ decreases as the distance between them increases. In the context of conversations in online forums, the distance between $prime$ and $target$ is difficult to define, because a long distance between two posts, whether it is calculated by time or by number of posts between them, does not necessarily result in the weak priming effect. Based on these considerations, we only compute RepDecay in the SWBD corpus, which solely consists of two-party dialogues. BNC corpus is also excluded because it contains multi-party dialogues that makes it difficult to extract a clear $prime$-$target$ relationship.

4.5 Propensity of alignment

We use all the posts/utterances produced by one individual to measure his/her propensity of alignment. For LLA and SCC, we use all of the ($prime$, $target$) pairs within a certain distance where individual I_i produces the $target$ to represent I_i's propensity of alignment. For RepDecay, we compute the regression coefficient β_i from the sequence of ($prime$, $target$) pairs in which $target$ is produced by I_i and use β_i to represent I_i's propensity of alignment.

We select only those active individuals from the four corpora whose number of posts/utterances is above a common threshold (above 90% of the population). For CSN corpus, we select 1066 active users who have composed at least 50 posts. For MOOC corpus, we select 829 active users who have composed at least 10 posts. For SWBD corpus, all 1296 speakers are selected. For BNC corpus, 502 active speakers who have at least over 26 utterances are se-

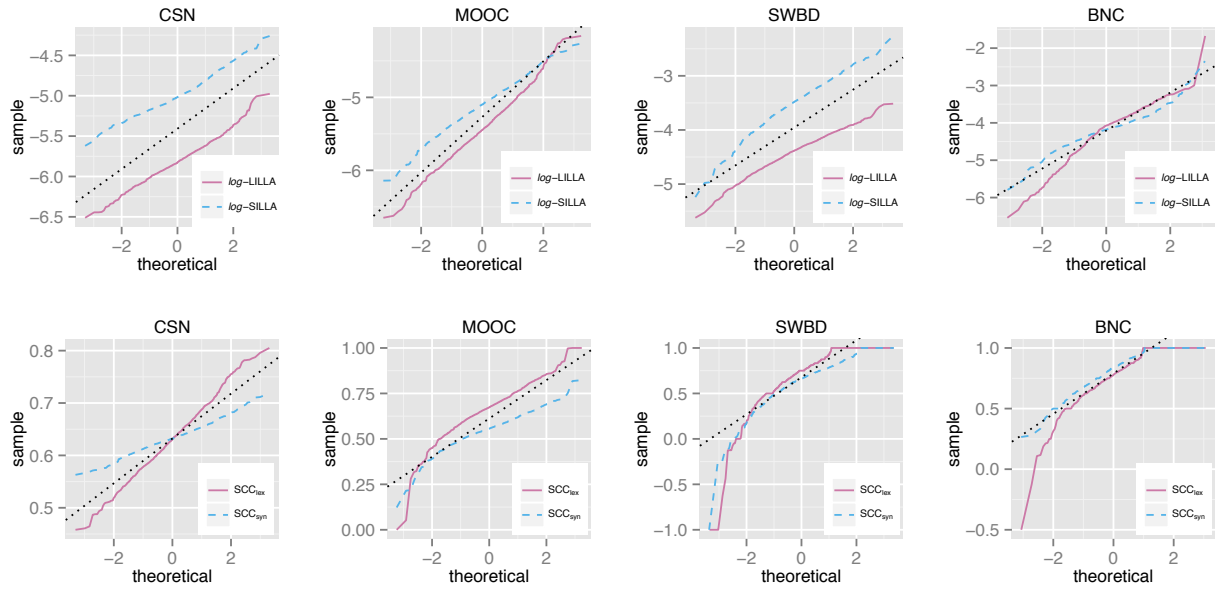

Figure 1: The quantile-quantile plots of LLA and SCC

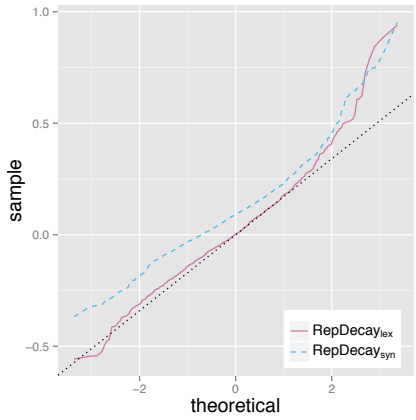

Figure 2: The quantile-quantile plot of RepDecay in SWBD

lected. These active forum-users or speakers are referred to as active individuals.

5 Intrinsic Evaluation Results

5.1 Normality of distribution

We use Shapiro-Wilt test (Shapiro and Wilk, 1965) to examine the normality of distributions of LLA and SCC in all of the four corpora, and the normality of distribution of RepDecay in the SWBD corpus (because RepDecay is only computed in SWBD). The test results show that all these distributions are significantly different from a normal distribution ($p < 0.001$).

But we can still use the quantile-quantile plot of each distribution to compare their normality relatively. Figure 1 show quantile-quantile plots of LLA and SCC in all of the four corpora, and Figure 2 shows the quantile-quantile plot of RepDecay in the SWBD corpus. It can be seen that the quantile-quantile plots of LLA and RepDecay are closer to straight lines (demonstrated by the dot-line) than SCC, thus they have relatively better normality in their distributions.

5.2 Sensitivity

We use NPS Chat Corpus (Forsyth and Martell, 2007) to construct several pieces of pseudo text with different levels of alignment strength, and then investigate the performance of the measures in revealing the difference.

The structure of the pseudo text assembles a sequence of turn-by-turn utterances in a dialogue. We control the strength of alignment by adjusting the probability of a word appearing in an utterance given whether it has appeared in the previous utterance or not. In a non-alignment control condition, the probability of the occurrence of a word is independent of

Table 2: Correlation coefficients between lexical and syntactic measures.

Measure	**CSN**	**MOOC**	**SWBD**	**BNC**
log-LILLA and log-SILLA	0.374***	0.237***	0.188***	0.369***
SCC_{lex} and SCC_{syn}	0.045***	-0.008	-0.001	0.200***
$RepDecay_{lex}$ and $RepDecay_{syn}$	NA	NA	0.695***	NA

$^{*}p < 0.05$, $^{***}p < 0.001$

Table 1: t-test results of comparing measures between different α values

$\alpha = 1$	t-score	
vs.	***log*-LILLA**	**SCC_{lex}**
$\alpha = 1.05$	-1.610	0.000
$\alpha = 1.10$	-2.704*	-0.061
$\alpha = 1.15$	-3.925**	-0.152
$\alpha = 2.25$	-17.47***	-2.463*
...	...	...
$\alpha = 3.00$	-22.23***	-2.839*

$^{*}p < 0.05$, $^{**}p < .01$, $^{***}p < .001$

its occurrence in the previous utterance. In conditions where alignment exists, this probability is dependent on the word's previous occurrences. For example, the prior probability of word "like" is 0.005, if it appears in the first utterance, then we set its probability to appear again in the second utterance is $0.005 * \alpha$ ($\alpha >= 1$), which is slightly larger than the prior. Larger α indicates higher strength of alignment between utterances, and $\alpha = 1$ indicates no alignment.

We use $\alpha = 1, 1.05, 1.1, ..., 3$, to construct sequences of text. Each sequence has 100 utterances, and each utterance randomly has 50 to 100 words. In each sequence, we compute the log-LILLA and SCC_{lex} measures for all the 99 pairs of adjacent pairs of utterances, i.e., *u1* and *u2*, *u2* and *u3* etc., using the precedent utterance as $prime$ and the following one as $target$. Finally we conduct pairwise t-test on the measures between the condition of $\alpha = 1$ and the conditions of other α values respectively (Table 1). RepDecay is not included in this analysis, because the decay effect is not considered when we construct the pseudo text.

Table 1 shows that LLA can detect the alignment effect at $\alpha = 1.10$ (at $p < 0.05$), while SCC can only detect $\alpha >= 2.25$. Thus, LLA has higher sensitivity than SCC.

6 Extrinsic Evaluation Results

As introduced in Section 3, we evaluate the performance of LLA, SSC, and RepDecay in three aspects: Consistency across different linguistic representation levels, between-individual difference, and within-individual stability.

6.1 Consistency

We calculate the Pearson correlation coefficients between lexical syntactic measures for LLA, SCC, and RepDecay (Table 2).

It is shown that the correlation between $RepDecay_{lex}$ and $RepDecay_{syn}$ is strongest, followed by the correlation between log-LILLA and log-SILLA. The correlation between SCC_{lex} and SCC_{syn} is only significant in CSN and BNC, but not in MOOC and SWBD. Thus, it indicates that RepDecay and LLA show better consistency between lexical and syntactic alignment than SCC.

6.2 Between-individual differences

We use one-way ANOVA to examine whether the between-individual differences of alignment propensity outweigh within-individual variance (Table 3). RepDecay is not included in the analysis because it generates only one value for each individual.

While all F scores indicate significant differences, LLA shows higher F scores than SCC. This result indicates that the alignment measures from some individuals are significantly higher than the others, and this tendency holds for both lexical alignment and syntactic alignment.

Table 3: F scores resulting from one-way ANOVAs (All values are significant at $p < 0.001$ level)

Measure	CSN	MOOC	SWBD	BNC
log-LILLA	15.05	2.761	51.52	14.32
log-SILLA	20.66	2.402	25.44	5.289
SCC_{lex}	8.884	1.205	3.448	1.937
SCC_{syn}	1.494	1.185	4.242	3.492

6.3 Within-individual stability

We use the coefficient of variation (CV) (Abdi, 2010) (also known as relative standard deviation), to evaluate the within-individual stability of the measures. CV is defined as the ratio of the standard deviation σ to the mean μ: $c_v = \sigma/\mu$. A smaller CV indicates less variability of a random variable in relation to its mean.

We calculate the CVs of LLA and SCC for each active individual in the four corpora, and then use t-tests to compare LLA vs. SCC (for lexical and syntactic measures respectively). RepDecay is not included in this analysis because it generates one value for each individual and thus there is no within-individual variance.

The t-tests results indicate that log-LILLA has smaller CVs than SCC_{lex} across the four corpora ($p < 0.001$). log-SILLA also has smaller CVs than SCC_{syn} for CSN, MOOC and SWBD corpora ($p < 0.001$), and there is no significant difference for BNC corpus ($p = 0.299$). This indicates that LLA has better within-individual stability than SCC.

7 Conclusions and Discussion

In this study, we evaluate the intrinsic properties of three computational measures of linguistic alignment: indiscriminate local linguistic alignment (LLA), Spearman's correlation coefficient (SCC), and repetition decay (RepDecay). We also evaluate their performance when applied to an extrinsic study about the IAM theory and individuals' alignment propensity.

From the intrinsic evaluations, we find that LLA and RepDecay are more normally distributed than SCC, and that LLA is more sensitivity than SCC. The main cause for the poorer normality of SCC roots in its way of computation: there has to be at least two common elements in order to get a valid value, but if $target$ is a pure repetition of $prime$, the value is always 1. Thus for short utterances that are common in spoken dialogues (SWBD and BNC), they are more likely to generate 1s, which result in the skewed distribution of SCC.

From the extrinsic evaluations, our main conclusions are: First, in terms of the propensity of alignment, both LLA and SCC can reveal significant individual differences. Meanwhile, LLA shows larger effect size for individual differences, and higher within-individual stability than SCC. Second, in terms of the correlation between alignment at the lexical and syntactic levels, RepDecay shows the strongest correlation, but LLA also consistently shows strong correlation across all corpora investigated. However, SCC does not consistently show this correlation.

Our study provides potential suggestions to future computational investigations about linguistic alignment. LLA is more favorable if the research question relates to individuals' inherent propensity of alignment, because it yields more significant between-individual differences and has better within-individual stability. LLA has better normality and sensitivity properties. RepDecay is more favorable if the research question is to explore the correlations between alignment at different linguistic levels, because it shows strongest correlation between lexical and syntactic levels in this study.

For future work, to explore the application of computational measures in revealing individuals' propensity of alignment at multiple linguistic levels (other than lexical and syntactic) could be an interesting direction.

Acknowledgments

We would like to thank the American Cancer Society (Kenneth Portier and Greta E. Greer) for curating and providing the Cancer Survivor Network corpus, and Anna Divinksy and Bart Pursel for the Art MOOC corpus. We thank John Yen and Yafei Wang for their helpful discussions.

References

Abdi, Hervé (2010). "Coefficient of variation". In: *Encyclopedia of Research Design. SAGE, Thousand Oaks, CA*, pp. 169–171.

BNC (2007). *The British National Corpus, version 3 (BNC XML Edition)*. URL: `http://www.natcorp.ox.ac.uk/`.

Branigan, Holly P., Martin J. Pickering, and Alexandra A. Cleland (1999). "Syntactic priming in language production: Evidence for rapid decay". In: *Psychonomic Bulletin and Review* 6.4, pp. 635–640.

Church, Kenneth W (2000). "Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p 2". In: *Proceedings of the 18th Conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 180–186.

Danescu-Niculescu-Mizil, Cristian and Lillian Lee (2011). "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs". In: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, pp. 76–87.

De Marneffe, Marie-Catherine, Bill MacCartney, Christopher D Manning, et al. (2006). "Generating typed dependency parses from phrase structure parses". In: *Proceedings of LREC*. Vol. 6, pp. 449–454.

Dubey, Amit, Patrick Sturt, and Frank Keller (2005). "Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling". In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 827–834.

Forsyth, Eric N and Craig H Martell (2007). "Lexical and discourse analysis of online chat dialog". In: *Semantic Computing, 2007. ICSC 2007. International Conference on*. IEEE, pp. 19–26.

Fusaroli, Riccardo et al. (2012). "Coming to terms quantifying the benefits of linguistic coordination". In: *Psychological Science* 8, pp. 931–939.

Garrod, Simon and Anthony Anderson (1987). "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination". In: *Cognition* 27.2, pp. 181–218.

Giles, Howard (2008). *Communication accommodation theory*. Sage.

Gnisci, Augusto (2005). "Sequential strategies of accommodation: A new method in courtroom". In: *British Journal of Social Psychology* 44.4, pp. 621–643.

Gries, Stefan Th. (2005). "Syntactic priming: A corpus-based approach". In: *Journal of Psycholinguistic Research* 34.4, pp. 365–399.

Huffaker, David et al. (2006). "Computational measures for language similarity across time in online communities". In: *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*. Association for Computational Linguistics, pp. 15–22.

Jones, Elizabeth et al. (1999). "Strategies of accommodation: Development of a coding system for conversational interaction". In: *Journal of Language and Social Psychology* 18.2, pp. 123–151.

Jones, Simon et al. (2014). "Finding Zelig in Text: A Measure for Normalizing Linguistic Accommodation". In: *25th International Conference on Computational Linguistics*. University of Bath.

Kilgarriff, Adam (2001). "Comparing corpora". In: *International journal of corpus linguistics* 6.1, pp. 97–133.

Marcus, Mitchell et al. (1994). "The Penn Treebank: annotating predicate argument structure". In: *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pp. 114–119.

Myers, Jerome L, Arnold Well, and Robert Frederick Lorch (2010). *Research design and statistical analysis*. Routledge.

Pickering, Martin J. and Simon Garrod (2004). "Toward a mechanistic psychology of dialogue". In: *Behavioral and brain sciences* 27.02, pp. 169–190.

Reitter, David, 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*. Association for Computational Linguistics, pp. 121–124.

Shapiro, Samuel Sanford and Martin B Wilk (1965). "An analysis of variance test for normality (complete samples)". In: *Biometrika*, pp. 591–611.

Wang, Yafei, David Reitter, and John Yen (2014). "Linguistic Adaptation in Conversation Threads: Analyzing Alignment in Online Health Communities". In: *Proc. Cognitive Modeling and Computational Linguistics. Workshop at the Mtg. of the Association for Computational Linguistics*.

Willemyns, Michael et al. (1997). "Accent Accommodation in the Job Interview Impact of Interviewer Accent and Gender". In: *Journal of Language and Social Psychology* 16.1, pp. 3–22.