Information density converges in dialogue: Towards an information-theoretic model

Yang Xu, David Reitter*

The Pennsylvania State University, USA

Abstract

The principle of entropy rate constancy (ERC) states that language users distribute information such that words tend to be equally predictable given previous contexts. We examine the applicability of this principle to spoken dialogue, as previous findings primarily rest on written text. The study takes into account the joint-activity nature of dialogue and the topic shift mechanisms that are different from monologue. It examines how the information contributions from the two dialogue partners interactively evolve as the discourse develops. The increase of local sentence-level information density (predicted by ERC) is shown to apply to dialogue overall. However, when the different roles of interlocutors in introducing new topics are identified, their contribution in information content displays a new converging pattern. We draw explanations to this pattern from multiple perspectives: Casting dialogue as an information exchange system would mean that the pattern is the result of two interlocutors maintaining their own context rather than sharing one. Second, we present some empirical evidence that a model of Interactive Alignment may include information density to explain the effect. Third, we argue that building common ground is a process analogous to information convergence. Thus, we put forward an information-theoretic view of dialogue, under which some existing theories of human dialogue may eventually be unified.

Keywords: dialogue, grounding, interactive alignment, uniform information density, entropy, information

1. Introduction

Dialogue can be understood from multiple theoretical perspectives. It is a *joint activity* in which language plays a prominent role, a dynamic process in which the common goal and the mutual understanding between speakers are achieved through *grounding* (Clark, 1996). It can also be viewed as a two-way

*David Reitter
Email address: reitter@psu.edu (David Reitter)

communication system where information flow follows general regularities, e.g., the rule of optimizing the rate of information transmission (Shannon, 1948; Genzel & Charniak, 2002, 2003), and the tendency to distribute material such that the density of information remains constant (*Uniform Information Density* hypothesis, UID, Jaeger & Levy 2006; Jaeger 2010; Temperley & Gildea 2015). Dialogue is also interpreted as a process in which alignment between interlocutors occurs at multiple levels of linguistic representation, as a result of primitive priming mechanisms (a.k.a, the *Interactive Alignment Model*, IAM) (Pickering & Garrod, 2004), which can result in reduced surprisal in terms of the language that each dialogue partner can observe. Whether alignment should be seen in an integrated theory as the cause, an epiphenomenon, or an additional process, is unclear; however, in this paper we explore the possibility of an account of dialogue that integrates these theories through observable, regular patterns of information distribution.

The motivation of this paper comes from two directions. First, in the work on the entropy rate constancy (ERC) principle and later in the UID framework, most of the existing studies rely on empirical evidence from corpora of written-form natural language. Only some preliminary work exists on transcripts of dialogue (Vega & Ward, 2009). It is unknown whether the inherent differences between dialogue and written text will bring up issues concerning the applicability of the theories. For example, questions can be asked such as "does the ERC/UID principle apply to the language from individual speaker alone, or to the dyad of interlocutors as a system where multiple inputs of information merge?"

Second, the other two theories mentioned above, IAM and grounding, have rich connections, but are not fully reconciled in terms of their subjects and scopes. For example, Pickering & Garrod's (2004) IAM proposes that the alignment of *situation models*, i.e., the multi-dimensional representation of the situation under discussion, is central to a successful dialogue, and that interlocutors proceed towards this higher level alignment by aligning on what they termed an *implicit common ground*. This concept analogizes to the notion of *common ground* by Clark & Brennan (1991), but whether they refer to the same process remains debatable.

Therefore, by answering the call for a better examination on how the information centered theories (ERC and UID) apply to the language production in dialogue, we propose a novel information-theoretic perspective of dialogue, which captures the basic needs of successful communication – effective and efficient information exchange, and also potentially reconciles the IAM and grounding theory. From a hierarchical view of language use in dialogue, the IAM covers a full range of linguistic representation levels, but most of the empirical work only examines alignment at lower levels, such as phonemes (Pardo, 2006), lexicon (Garrod & Anderson, 1987) and syntactic structures (Pickering & Branigan, 1998; Branigan et al., 2000; Reitter & Moore, 2014). On the other hand, grounding theory (Clark, 1996) provides mostly qualitative descriptions at higher levels by viewing dialogue as indistinctive example of joint activities in general sense. The approach taken in this study investigates dialogue at a level somewhere be-

tween the IAM and grounding perspectives, which starts with quantifying the information density at sentence level, and then takes into account the unique discourse structure in spoken conversations.

The first step of our work is to quantify the sentence information (The reason for focusing on sentence is explained in Section 2.1) in dialogue, and to examine whether it demonstrates the same overall increasing pattern that has been discovered in written language. The purpose of this step is to confirm whether the shared context between interlocutors consistently accumulates in dialogue. In the second step, we zoom in to the level of the topic episodes (The reason of doing so is discussed in Section 2.2), which are delineated using a topic segmentation technique. We focus on the information per word changes near the boundaries of topic episodes. The goal is to relate the information interchange between interlocutors to the process of building common ground, as well as the alignment approach. Third, we turn to alignment, hypothesizing that it is either cause or consequence of the informational exchange that requires coordination between speakers. We observe levels of linguistic alignment within topic episodes.

The remainder of this paper is organized into seven sections. After an introduction of background and raising the research questions in Section 1, some previous studies of related issues are reviewed in Section 2. In Section 3 we examine the overall trend of information density in dialogues. In Section 4 and Section 5, we examine the effect of topic shifts on sentence information, and demonstrate the information converging patterns between interlocutors of different roles within the scope of topic episodes. Then in Section 6, we demonstrate how lexical alignment can be used to explain the convergence pattern. Finally, the implications of these observations are discussed in Section 7.

2. Related Work

2.1. The principle of entropy rate constancy

Human communication systems such as written text and speech have been claimed to be optimized in that the rate of information being transmitted keeps constant and is close to the channel capacity, a.k.a., following principle of *entropy rate constancy* (ERC) (Genzel & Charniak, 2002, 2003; Qian & Jaeger, 2011). This line of work was inspired by Information Theory (Shannon, 1948), which relates uncertainty about the next signal to the amount of information that can be transmitted. The observation that communicators tend to distribute information evenly may have much to do with the idea that information is a measure of cognitive load: much information, or at least much surprisal at the information conveyed, is a model of how difficult it is to process (Hale, 2001).

In Genzel & Charniak's (2002) original work, the amount of information conveyed in natural language is estimated by treating the word as a random variable, and then a bulk of text becomes a sequence of random variables X_i, where X_i corresponds to the i-th word in the text. ERC predicts that the information amount of the conditional random variable, $X_i|X_1 = w_1, \ldots, X_{i-1} =$

w_{i-1}, should remain constant as i increases. Because natural language organizes messages into sentences, the context formed by all previous words $X_1 = w_1, \ldots, X_{i-1} = w_{i-1}$ can be decomposed into two parts: The *global context* C_i, all the words from preceding sentences, and the *local context* L_i, all the preceding words within the same sentence as X_i. Then the variable that remains constant can be written as the left term in Equation 1, where H refers to entropy. According to Information Theory, this term can be further decomposed into the two terms on the right side: $H(X_i|L_i)$, the entropy of X_i conditioned on local context L_i, and $I(X_i, C_i|L_i)$, the conditioned mutual information between X_i and its global context C_i.

$$H(X_i|C_i, L_i) = H(X_i|L_i) - I(X_i, C_i|L_i) \tag{1}$$

Intuitively, Equation 1 says that knowing about the global context (i.e., having positive mutual information between X_i and C_i), will make the words under the current local context more predictable (hence, having lower entropy). Another important fact is that as i increases, the mutual information term $I(X_i, C_i|L_i)$ will also increase, because knowing more about the context makes it easier to predict the upcoming content. Since the whole right side of Equation 1 also needs to remain constant (according to Information Theory), then the entropy conditioned on local context, $H(X_i|L_i)$ must also increase with i. Therefore, the increase of locally-conditioned entropy $H(X_i|L_i)$ is an indicator of the ERC principle in natural language.

Theoretically, the entropy of a random variable is defined as the expected value of the information conveyed, and the amount of information is measured by the negative logarithm of the probability of the event (Shannon, 1948). For a discrete random variable X, which has n possible outcomes, its entropy is:

$$E[-\log(P(X))] = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2}$$

where the unit of entropy is *bit* when the base 2 logarithm is computed.

Applying these concepts to natural language, the random variable of our interest here is the next word given its preceding context (words), a.k.a., an N-gram. For example, to estimate the entropy of the third word X after the bigram context *this is*, then ideally, we need to enumerate all possible words that follow the context, e.g., *good*, *bad*, *great*, etc., and estimate their probabilities, $P(good|this\ is)$, $P(bad|this\ is)$, $P(great|this\ is)$. The conditional entropy of X, $E[P(X|this\ is)]$, is computed in the same way as Equation 2. However, this method is impractical because it is nearly impossible to enumerate all the possible outcomes of X and estimate their probabilities accurately.

Therefore, alternative methods need to be used to properly estimate the entropy of words. Genzel & Charniak (2002) provide such a method by averaging the negative logarithm of probabilities of all the trigrams in a sentence, and use this per-word information to approximate the average entropy of words in the sentence (see Section 3.1 for details). Thus, Genzel & Charniak (2002) compute

the $H(X_i|L_i)$ term in Equation 1 at the sentence level, and they have confirmed that this variable increases with i, which now indicates the position of a sentence within the text.

We adopt Genzel & Charniak's method in this study and examines how the estimated information of sentence is distributed and evolves in dialogue. Genzel & Charniak (2002, 2003) in their original work and Keller (2004) in a following study all referred to the per-word average information as *entropy*, although this measure does not follow the definition of Shannon's entropy (Equation 2). Because *entropy* is a measure of information that can be potentially conveyed over a channel before seeing the actual transmission, as opposed to the actual *information* conveyed by actual signals, which Genzel & Charniak's approximation quantifies. Through the remainder part of this paper, we use the term *sentence information* instead of entropy to account for the post-hoc nature of the approximation of $H(X_i|L_i)$.

More recently, the idea of *uniform information density* (UID, e.g., Jaeger & Levy 2006) has extended ERC into a broader framework that governs how people manage the amount of information in language production, from lexical levels to all levels of linguistic representations, e.g., syntax or semantics. The core idea of UID is that people avoid salient changes in the density of information (i.e., amount of information per amount of linguistic signal) by making specific linguistic choices under certain contexts (Jaeger, 2010). Therefore, UID could be viewed as a generalization of the principle of ERC.

We now take this a step further. We postulate that ERC/UID is a principle that applies not merely to an individual's cognition, but to the system formed by several interlocutors as a whole. In examining information density in dialogue we will point out that information density in an individual's language (not necessarily their cognitive processes) does not always follow the principle if they lead rather than follow the conversation, but that the system of two speakers displays a constant entropy rate.

While there is a body of literature on ERC and UID, spoken conversation is rarely examined, and when it is, conversation (like Twitter messages) is treated as if it was monologue. Yet, authors widely acknowledge that speech in dialogue is different from written language in both form and content. For example dialogue is different from monologue in that it is inherently interactive and contextualized, which results in its "irregular grammaticality" and "theoretically uninterested complexities that are unwanted" (Pickering & Garrod, 2004). Considering the differences between spoken and written language, we realize that many previous studies on human communication that use information measurement method are incomplete in methodology and not comprehensive enough in conclusions as well, because they solely use written language as experiment materials. Thus, to better understand human communication from the perspective of information theory, a careful investigation is necessary on how much information each party contributes in a dialogue and how the proportion of contribution develops.

2.2. Sentence information and discourse structure

The ERC principle also leads to an interesting prediction about the relationship between information change and topic shift in text. Generally, a sentence that initiates a shift in topic will have lower mutual information with its context, because the preceding context provides little information to the new topic. Thus, a topic shift corresponds to a drop in the mutual information term $I(X_i, C_i|L_i)$ in Equation 1. Then, to keep the left term constant, as predicted by ERC, the local information term, $H(X_i|L_i)$, needs to decrease when a topic shift happens. Genzel & Charniak (2003) verified this prediction by showing that paragraph-starting sentences have lower information content than non-paragraph-starting ones, with the assumption that a new paragraph often indicates a topic shift in written text. This paragraph effect does not exist consistently across different genres of text, because a new paragraph does not necessarily represent a new topic.

Following this line of work, Qian & Jaeger (2011) applied a more fine-grained latent topic modeling approach to ask how topic shifts affect sentence information. They found the expected negative correlation between sentence information and topic shift, and that the effect of topic shift subsumes the effect of sentence position (for details, see Qian & Jaeger, 2011). The relationship between sentence information and discourse structure has also been utilized to build topic segmentation tools (Eisenstein & Barzilay, 2008), in which the optimization of segmentation is equivalent to minimizing the weighted sum of entropies. More recently, Doyle & Frank (2015) leverage Twitter data about ongoing baseball games to find further support to the ERC principle: the information content of messages gradually increases as the context builds up, and it sharply goes down when there is a sudden change in the non-linguistic context.

The topic shift in spoken dialogue is not explicitly designated by paragraph structures. Rather, as a dialogue unfolds, topic changes naturally happen when a current topic is exhausted or a new one occurs (Ng & Bradac, 1993; Linell). In the field of *Conversation Analysis* (CA), the basic unit of topical structure analysis in dialogue is *episode*, which refers to a sequence of speech events that are "about" something specific in the world (Linell). To be precise, we use the term *topic episode* in this study.

The formation of a topic episode is a joint accomplishment of two speakers and a product of initiatives and responses (Linell, 1990). When establishing a new topic jointly, one speaker first produces an initiatory contribution that introduce a "candidate" topic, and the other speaker makes a response that shares his perspective on that (Linell). The *initiator* of a new topic introduces *novelty* or *surprisal* into the context, while the other speaker, the *responder*, is more of a *commenter* or *evaluator* of information, who does not contribute as much in terms of novelty. Therefore, the establishment of topic shift can be viewed as a joint activity between interlocutors. Clark (1996) models joint activities as sequences of *sub-activities*. If we view a dialogue as a joint activity by two participants, then the topic episodes that naturally form within can also be viewed as the so-called sub-activities. Furthermore, Clark (1996) points

out that participants have roles to play in a joint activity, which may change from sub-activity to the next. This argument directly supports the initiator vs. responder distinction in topic shift. The *initiator* and *responder* are such roles that help them coordinate to accomplish the goal of the sub-activity, i.e., creating a new topic.

Relating these theoretic insights back to the information approach of natural language, it is reasonable to anticipate the same effect of sentence information decrease patterns near the boundaries of topic episodes in dialogue. Considering the *initiator* vs. *responder* discrepancy in speaker roles, we also expect their patterns of change in information density to be different.

2.3. Grounding and IAM

Language-as-activity means that dialogue is a joint activity during which multiple (two or more) interlocutors contribute alternatively to *common ground* (Clark & Brennan, 1991). The notion of common ground can be traced back to Stalnaker (1978), based on older, similar concepts such as *common knowledge* (Lewis, 1969), *mutual knowledge or belief* (Schiffer, 1972), and *joint knowledge* (McCarthy, 1987). Clark (1996) summarizes the definition of common ground: the common ground between two interlocutors is the sum of their mutual, common, or joint knowledge, beliefs, and suppositions. To clarify the confusion about these notions, refer to Clark (1996), who also gives a detailed discussion on the different representations of common ground.

As mentioned in Section 1, Pickering & Garrod (2004) incorporated the concept of common ground into the framework of IAM, by treating *implicit common ground* as one of the representation layers that becomes aligned between interlocutors, which is the automatic consequence of the lower-level alignments. In their work, implicit common ground is understood as a "subset" of the original notion of common ground, which does not refer to the complete background knowledge shared between interlocutors (Clark & Marshall, 1981), but rather something dynamically built up along the course of dialogue from the alignment at lower levels. However, there is still debate over using the notion of common ground in IAM. For instance, Barr & Keysar (2004) explain that the term "common ground" in Clark & Marshall (1981) refers to the *meta*-knowledge rather than the *shared* knowledge, and that what Pickering & Garrod (2004) are referring to by "implicit common ground" is really just shared knowledge.

3. The Dynamics of Sentence Information in Dialogue

As the first step in building an information-theoretic account of dialogue, we examine whether the principle of ERC applies to spoken dialogue at all. We anticipate that sentence information should increase throughout each conversation just as it does in written text. This step serves as the basis for the rest of this study.

Table 1: Basic statistics of the corpora.

Statistics	**SWBD**	**BNC**
No. of dialogues	1126	1346
Avg no. of turns per dialogue	109.3 ($SD = 50.7$)	51.7 ($SD = 102.9$)
Avg no. of sentences per dialogue	141.0 ($SD = 61.4$)	70.3 ($SD = 133.9$)
Total no. of words	7.14 M	3.88 M

3.1. Method: Estimate sentence information

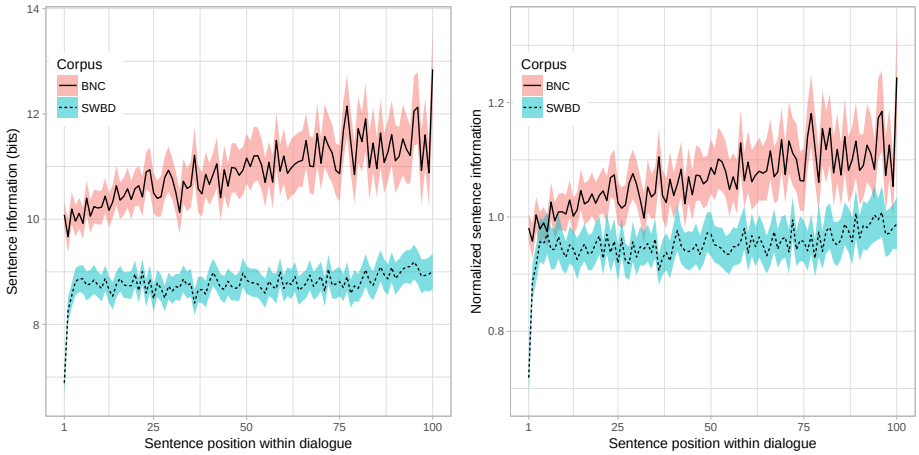
The Switchboard corpus (SWBD, Godfrey et al. 1992) and the British National Corpus (BNC, BNC 2007) are used. SWBD contains 1126 dialogues over telephone between two native American English speakers. The complete BNC dataset is a collection of written and spoken language of British English. Its spoken part 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 SWBD and to simplify our experiment design, we select part of BNC-DEM that contain two speakers within each conversation. For convenience, in the rest part of this paper, we use BNC to refer to this sampled part of BNC-DEM. Some basic statistics of the two corpora are shown in Table 1.

We use trigram language models (LMs) to estimate the information of a sentence, which is similar to Genzel & Charniak's (2003) method. A sentence is considered as a sequence of words, $S = \{w_1, w_2, \ldots, w_n\}$. The information of the whole sequence, $H(S)$, is estimated by averaging over the negative logarithms of all the trigram probabilities in the sentence:

$$\begin{aligned} H(S) = H(w_1 \ldots w_n) &\approx -\frac{1}{n} \sum\nolimits_{w_i \in W} \log P(w_i | w_1 \ldots w_{i-1}) \\ &\approx -\frac{1}{n} \sum\nolimits_{w_i \in W} \log P(w_i | w_{i-2} w_{i-1}) \end{aligned} \tag{3}$$

where $P(w_i|w_{i-2}w_{i-1})$ is estimated by the LM using Katz backoff (Katz, 1987). The software we use to train the LMs is SRILM (Stolcke, 2002). To estimate the probability terms as accurately as possible, LMs of higher performance (hence, lower perplexity, etc.) are preferred. Because the selection of training set matters to the performance of LMs, we carefully compared several options (see Appendix A), before we finally use the method as below.

We extract the first 100 sentences from each conversation (this number is about the average length of conversations), and apply a *position-wise* 10-fold cross-validation. Specifically, we randomly divide the data into 10 subsets, S_i ($i = 1, 2, \ldots, 10$), and in each round of cross-validation, we train LMs from $\{S_j | j \neq i\}$, and use them to compute the sentence information in S_i. "Position-wise" means that when computing the information of a sentence, we use LMs



(a) Sentence information (b) Normalized sentence information

Figure 1: Sentence information (a) and normalized sentence information (b) against the global position (ranging from 1 to 100). The y-axis values in (b) are given $+.05$ and $-.05$ offsets on BNC and SWBD respectively to avoid overlap. Bootstrapped 95% confidence bands.

trained from sentences of the same position, i.e., for the sentences of position k in S_i, S_i^k $(k = 1, 2, \ldots, 100)$, we need to use the LM trained from $\{S_j | j \neq i,\ \text{position} = k\}$. Therefore, in each cross-validation round, we train 100 distinct LMs. Genzel & Charniak (2003) first used this position-wise training method on the *Wall Street Journal* text, but they did not explain the reason behind it. Presumably, their intention was to avoid using information from the later part to predict the preceding content. Although we have found that the non-position-wise training method draws similar results to the position-wise training (see Appendix A), we adopt the latter method here so that the results are comparable to the previous work.

3.2. Results

3.2.1. Sentence information increases in dialogue

Figure 1a shows the sentence information against its *global position*, i.e., the sentence position from the beginning of dialogue (from 1 to 100). It can be seen that sentence information increases with global position in both corpora. BNC has larger slope, and SWBD has a flatter curve but sharper increase at the early stage of dialogue.

Tests of normality (see Appendix C) show that the distributions of sentence information in both corpora are significantly different from normal ones, and that its logarithm seems to have better normality. Thus, in the later part of this section where statistical tests are conducted, we take the logarithm transformation of sentence information as response variable. It is worth noting that the *log* transform of information does not convey theoretical meaning, but it is

just an intermediate variable that better fits the prerequisites of the statistical tests.

Based on the considerations above, in order to test the statistical significance of the observed increase of sentence information in Figure 1a, we fit linear mixed-effect models, using the logarithm of sentence information as response variable, and the global position as predictor (fixed effect), with a random intercept grouped by distinct dialogues. The `lme4` package in `R` is used (Bates et al., 2015). It shows reliable fixed effects of global position for both SWBD ($\beta = 3.9 \times 10^{-4}, p < 0.001$) and BNC ($\beta = 1.4 \times 10^{-3}, p < 0.001$).

We notice that the curve of SWBD looks flat after a boost at the early stage. We fit an extra (unplanned) model to test whether the increase is reliable in later stage of dialogue (for global position ≥ 10) as well. This is indeed the case ($\beta = 2.1 \times 10^{-4},\ p < 0.01$), although the regression coefficient (reflecting how fast sentence information increases) is smaller than that of BNC. The early dramatic increase of sentence information in SWBD, and the difference in regression coefficients between corpora might be related with the forms of conversation. The low-information sentences at the beginning of SWBD conversation could convey common greetings in conversations over the phone. On the other hand, BNC conversations took place face-to-face, and it is possible that the two participants had met before the conversation started, which resulted in the relatively high initial sentence information (because they started the substantial content right away). More quantitative and qualitative work is needed to investigate the between-corpora difference in the increase rate of sentence information, which is beyond the scope of this study.

3.2.2. Eliminating the effect of sentence length

Keller (2004) pointed out that in written text (e.g., *Wall Street Journal* articles) the increase of sentence information along global position could be an artifact of sentence length, i.e., sentence length increases with its position, and sentence information is known to be correlated with its length. He regresses out the effect of sentence length by computing the partial correlation (Johnson et al., 2014), and eventually shows that sentence information does increase with sentence position.

Here, our preliminary computation shows that in SWBD and BNC, sentence length and global position are weakly correlated (SWBD, $r = -0.035,\ p < 0.001$; BNC, $r = 0.038,\ p < 0.001$)[1], and sentence length is correlated with sentence information as well (SWBD, $r = 0.258,\ p < 0.001$; BNC, $r = 0.091,\ p < 0.001$). Therefore, it is necessary to eliminate the effect of sentence length.

Following the method proposed by Keller (2004), we compute the partial correlation coefficients between the total amount of information within sentence (i.e., without averaging in Equation 3) and its global position, with sentence

[1]Note that these two correlations are in opposite directions, which is probably caused by the different nature of the two corpora. Investigating this discrepancy may lead to some interesting findings, but it is beyond the scope of this study.

length partialled out. We find significant partial correlations for both SWBD ($r = 0.055,\ p < 0.001$) and BNC ($r = 0.117,\ p < 0.001$). This result is sufficient for us to conclude that sentence information does increase with global position when sentence length is controlled.

Considering that sentence information is correlated with sentence length, we hope to find a substitute of the original measure of sentence information, which on the one hand reflects the amount of information in a sentence, but is also independent of sentence length. We consider a method proposed by Genzel & Charniak (2003), which is originally used to calculate the syntactic complexity (*tree depth* and *branching factor*) of a sentence in a way that is independent of the sentence length.

Let $H(s)$ be the original information of a sentence s (computed via Equation 3), and $H'(s)$ be the target measure that is independent of sentence length. First, we need to compute $\bar{H}(n)$, the average per-word information of sentences of the same length n, for all possible lengths ($n = 1, 2, \dots$) that have occurred in the corpora:

$$\bar{H}(n) = \frac{1}{|S(n)|} \sum\nolimits_{s \in L(n)} H(s)$$

where and $S(n) = \{s | l(s) = n\}$ is the set of all the sentences in the corpus that have the length of n. Then we compute $H'(s)$ by:

$$H'(s) = \frac{H(s)}{\bar{H}(n)}$$

We refer to this new metric as the *normalized information*. It is not sensitive to the length of sentence. We still use the original information measure $H(s)$ to demonstrate some results in this paper, because it is a direct metric of the absolute magnitude of information.

How the normalized sentence information changes with global position is shown in Figure 1b. Basically, the normalized information increases with global position, which is consistent with the unnormalized sentence information. We fit linear mixed-effect models using the logarithm of the normalized information as the response variable, and global position as the predictor (fixed effect), with a random intercept grouped by dialogues. The models demonstrate significant fixed effects of global position for both SWBD ($\beta = 5.7 \times 10^{-4},\ p < 0.001$) and BNC ($\beta = 1.4 \times 10^{-3},\ p < 0.001$). We also fit an extra model for SWBD with global position ≥ 10, and the effect remains significant ($\beta = 3.0 \times 10^{-4},\ p < 0.001$).

Summarizing this section, we have found that sentence information increases over the course of the whole dialogue, which is consistent with previous findings on written text (Genzel & Charniak, 2002, 2003; Qian & Jaeger, 2011). This increase pattern remains significant when we eliminate the effect of sentence length. Our results lend further support to the ERC principle in the realm of speech communication, i.e., the way people organize information transmitting rate when they talk face-to-face or in phone is indeed governed by the generic

rule of efficient communication.

4. The Effect of Topic Shifts on Sentence Information in Dialogue

As we have seen, sentence information seems to change in ways that are similar in dialogue and in written text. However, the distribution of information becomes quite interesting once we account for the topic structure in dialogue. Previous studies have observed the change of sentence information caused by the topic shift in written text (Genzel & Charniak, 2003; Qian & Jaeger, 2011). Intuitively, this effect should also exist in dialogues, because the shift, maintenance and resume of topic episodes are ubiquitous in natural conversations (Ng & Bradac, 1993).

Empirically, the key is to identify the boundaries where topic shifts occur. Within the context of dialogue, this task involves finding, or defining, the *topic episodes* in dialogue. Here we compare two ways to accomplish the task. In a first approximation, we use speaking turns as topic episodes; and then, we apply a topic segmentation algorithm.

4.1. Speaking turns as topic episodes

It is simplistic to treat all the speaking turns as topic episodes, because a considerable proportion of spoken dialogues are short utterances that consist of *fillers* (e.g., "*uh*", "*um*", etc.), back-channeling (e.g., "*yeah*", "*mm*"), or incomplete sentences (see below for an example of short turns from SWBD). The semantic contribution of these turns depends largely on their context, and thus it is unreasonable to treat them as individual topic episodes.

A: Yeah.
B: Smaller towns.
A: Yeah. Smaller towns.
B: Oh.

However, for longer speaking turns that contain two sentences or more, it is sometimes reasonable to treat them as independent topic units, because they often convey relatively complex meanings. In the following example of long turns (partially shown) from SWBD, apparently the two speakers are expressing complex ideas, and each turn is about a relatively independent topic.

A: It was human nature. But it won't have can any, uh, any bad stuff. So, uh, I think I, we spend, of all the major semiconductor firms, we probably put safety and environmental on the utmost, foremost, uh, uh, first thing we always look at.
B: Well, I know from some of the sites that we've had, uh, quite a list of cites that have gone bad and you have to clean up. And, you know the law now is the super fund and anybody who's contributed toxic waste, no matter if you were somebody that eventually, you know, uh, damaged the ground or not. . . .

Based on the considerations above, we only treat the turns that consist of two sentences or more as topic episodes. Table 2 shows that about 20% turns in both corpora meet this mark. The remaining 80% turns that contain only one sentence are not included in the next analysis.

Table 2: Basic statistics of the turn length (in sentences)

	SWBD	**BNC**
Average turn length	1.29	1.54
Turn length standard deviation	0.70	2.63
Percentage of turn length = 1	79.5%	78.3%
Percentage of turn length = 2	15.0%	13.7%
Percentage of turn length = 3	3.7%	3.8%
Percentage of turn length > 3	1.8%	4.2%

On the basis of Genzel & Charniak's (2003) finding that sentence information drops at the beginning of new topics in written text, we expect to see the same pattern within a turn. Indeed, the normalized information of the non-turn-starting sentences is significantly higher than that of the turn-starting sentences (SWBD: $t = -6.29, p < 0.001$; BNC: $t = -14.51, p < 0.001$). This result implies that the long turns in dialogue function as individual topic units, which entails a distribution of information between sentences that is predicted by ERC.

4.2. Topic segmentation using TextTiling

The analysis in the previous section has obvious flaws. First, relatively long turns that contain more than two sentences make up only a small portion of each corpus (see Table 2). Second, if we consider the joint nature of the formation of new topics in dialogue (Linell, 1990), turns really make a poor man's approximation of a topic model. Therefore, we use topic segmentation algorithms to find topic shifts.

4.2.1. Method

There are multiple computational algorithms for the task of topic segmentation, such as the TextTiling algorithm (Hearst, 1997), Bayesian model (Eisenstein & Barzilay, 2008), Hidden Markov model (Blei & Moreno, 2001), graph-based model (Malioutov & Barzilay, 2006) etc. We have obtained similar results for our experiment (will be explained later) using TextTiling and other two state-of-the-art segmentation algorithms, *BayesianSeg*, a Bayesian unsupervised topic segmentation method (Eisenstein & Barzilay, 2008), and `MinCutSeg` a graph-based segmentation model (Malioutov & Barzilay, 2006). However, we find TextTiling to be a better option and will use it in the following demonstration, for two considerations: First, it outputs more reasonable segment length (see Appendix B for details). Second, it is a cohesion-based method, which avoids the confounding effect in more sophisticated methods that use word surprisal per se.

Here is the experiment procedure: We use the TextTiling algorithm to insert boundaries into the sequence of all sentences in a dialogue. Then we treat the resulting segment between any two boundaries as a *topic episode*, which is the

basic topic structure in dialogue. To give an impression of the performance of TextTiling on our data, we show two examples of the surrounding text near the topic boundaries:

Example 1 (from SWBD):

Prev episode	Speaker A: *So,* ***I was very comfortable****, you know, in doing it when it got to the point that we had to do it.*
Next episode	Speaker A: *But there's, well, I had an occasion for my mother-in-law who had fell and needed to be, you know, could not take care of herself anymore, was confined to a nursing home for a while that was* ***really not a very good experience****.*

Example 2 (from BNC):

Prev episode	Speaker B: *We need the keyboard now. I gotta find out where that plugs in at the back somewhere usually.*
Next episode	Speaker B: *Then we can* ***turn it on*** *and see if it all works. Can you hold that for a while?*

In the examples shown above, the horizontal lines indicate the topic boundaries. In the "previous episode" of Example 1, the two speakers were talking about "sending family members to nursing homes" and their attitude were basically positive (see the bold text). In the "next episode", however, speaker *A* brought up a counterexample that was "not a very good experience" (bold part). In Example 2, the two speakers were assembling a machine in the "previous episode", and their talks were basically about finding parts of the machine and trying to put them together. In the "next episode", they entered a new stage where their goal was to turn on the machine and test whether it functioned well. From the two examples, we can see that the TextTiling algorithm can indeed capture the topic episodes that naturally shape during dialogue, which are either caused by an active shift of topic, or the starting of new sub-activities.

For each topic episode, an *episode index* is assigned, indicating the episode's relative position in the dialogue (starting with 1). An additional *within-episode position* indicates a sentence's relative position from the beginning of the episode that it belongs to. Table 3 shows the average number of episodes per dialogue and the average number of sentences per episode in the two corpora.

4.2.2. Results

We plot the sentence information and normalized sentence information against the within-episode position of sentences, grouped by the episode index (Figure 2). Limited by space, we only show the first six topic episodes in each dialogue and the first ten sentences in each episode. It can be seen that sentence information and normalized information are of lower values at the beginning of topic episode, and they increase within the episode.

Table 3: Number of topic episodes per dialogue and number of sentences per episode resulted from applying the TextTiling algorithm.

Statistics	**SWBD**	**BNC**
Avg. no. of episodes per dialogue	12.1 ($SD = 4.3$)	7.1 ($SD = 10.8$)
Avg. no. of sentences per episode	11.7 ($SD = 10.6$)	12.4 ($SD = 8.3$)

Linear fixed-effects models were fitted to the data using the sentence information and normalized sentence information as response variable respectively, and the within-episode position as predictor (fixed effect), with a random intercept grouped by distinct episodes. We find a significant effect of within-episode position on both measures in both corpora: Sentence information in SWBD, $\beta = 5.9 \times 10^{-4},\ p < 0.001$; normalized information in SWBD, $\beta = 4.5 \times 10^{-3},\ p < 0.001$; sentence information in BNC, $\beta = 2.5 \times 10^{-2},\ p < 0.001$; normalized information in BNC, $\beta = 3.0 \times 10^{-3},\ p < 0.001$.

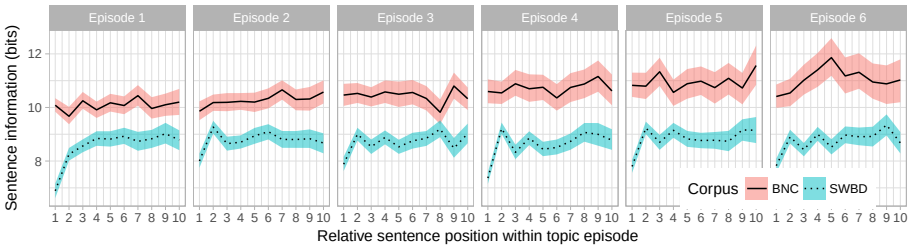
In summary, we have shown that by applying topic segmentation algorithm we can capture the topic structure in dialogue, i.e., *topic episodes*. The patterns of sentence information within the topic episodes and near their boundaries are consistent with what have been previously found on written text.

5. Speaker Roles in Dialogue Topic Segments

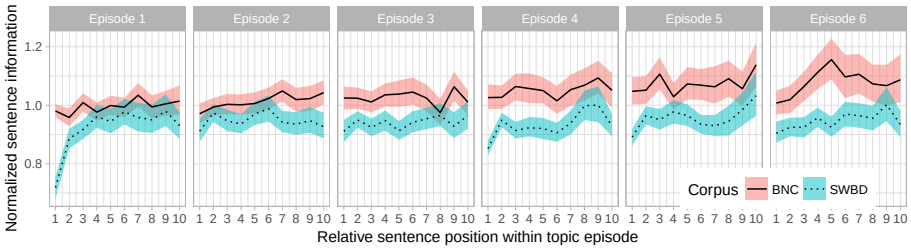
So far, we have characterized how the sentence information changes within the whole course of dialogue, and the patterns that are associated with the topic structure of dialogue. The previous two sections set the stage for our core purpose in this study, i.e., to model the joint and individual contributions of interlocutors from an information-theoretical perspective.

One feature of the "joint activity" nature of dialogue is that all the sentences are from two (in the simplest case) different interlocutors. So, a logical starting point would be to examine the sentence information of different interlocutors respectively. This requires us to systematically distinguish the two interlocutors. Linking back to Section 2.2, the unique topic shift mechanism in dialogue provides a rule to differentiate the speakers by categorizing them as either *topic initiator* or *topic responder* within the scope of each topic segment. The initiator of a new topic is the one who brings "novelty" to the current context, and a responder is the one who passively accepts or comments on the topic shift. We expect that this discrepancy in speaker roles can be reflected in their respective sentence information patterns. Examining that also lets us address the question whether and how ERC (and related UID) applies to an individual's language output, or whether it is a property of the dialogue partners as a communicating and collaborating system.

In the following part of this section, we operationalize the concepts of *topic initiator* and *topic responder*. Based on that we further investigate how their



(a) Sentence information



(b) Normalized sentence information

Figure 2: Sentence information (a) and normalized sentence information (b) against the relative sentence position (from 1 to 10) within topic episode grouped by topic index (from 1 to 6). The y-axis values in (b) are given $+.05$ and $-.05$ offsets on BNC and SWBD respectively to avoid overlap. Bootstrapped 95% confidence bands.

sentence information develop within the scope of topic episodes in dialogue.

5.1. Method: Identify two types of topic shifts

Topic shifts can be characterized according to whether they occur at turn boundaries or not. So, we identify *within-turn* topic shifts, which occur in the middle of a turn, and *between-turn* topic shift at the gap between two turns from two different speakers. In SWBD, 27.2% of the topic boundaries are within turns, and 72.8% are between turns, and for BNC the percentages are 41.2% and 58.8% respectively.

A within-turn topic boundary suggests that the speaker of the current turn is *initiating* the topic shift. On the other hand, a between-turn boundary suggests that the following speaker who first gives a substantial contribution is more likely to be the initiator of the new topic. Now that the topic initiator is identified, we can also mark the *topic initiating utterance* (TIU), to refer to the body of sentences that brings up the new candidate topic by the initiator.

We operationalize TIU as follows: for within-turn boundaries, let TIU be the remaining part of current turn after the boundary; for between-turn boundaries, let TIU be the whole body of the next turn immediately following the boundary. One empirical characteristic of a TIU is that it tends to be relatively long, because a short sentence is less likely to convey adequate information about a new topic. Thus, we validate this operational definition above by examining the

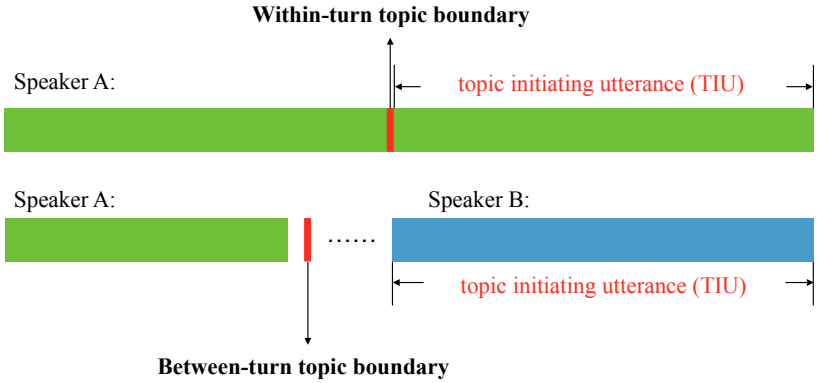


Figure 3: Operational definition of topic initiating utterances (TIUs). The red vertical bars indicate the topic boundaries placed using TextTiling. A complete horizontal bar of one color represents a turn from one speaker (green for speaker A and blue for speaker B). The upper line shows the case of within-turn topic boundary, and the lower line shows the case of between-turn topic boundary.

length (number of words) of TIUs. It shows that for within-turn boundaries, TIUs are relatively long (SWBD, $mean = 25.5$, $median = 19.0$; BNC, $mean = 25.3$, $median = 15.0$) as expected. But for between-turn boundaries, TIUs are short (SWBD, $mean = 9.3$, $median = 2.0$; BNC, $mean = 9.7$, $median = 5.0$). It means the definition of TIU for the case of between-turn boundaries needs to be modified, so that a more suitable portion of utterances are selected, and also an equitable selection is made between within- and between turn boundaries.

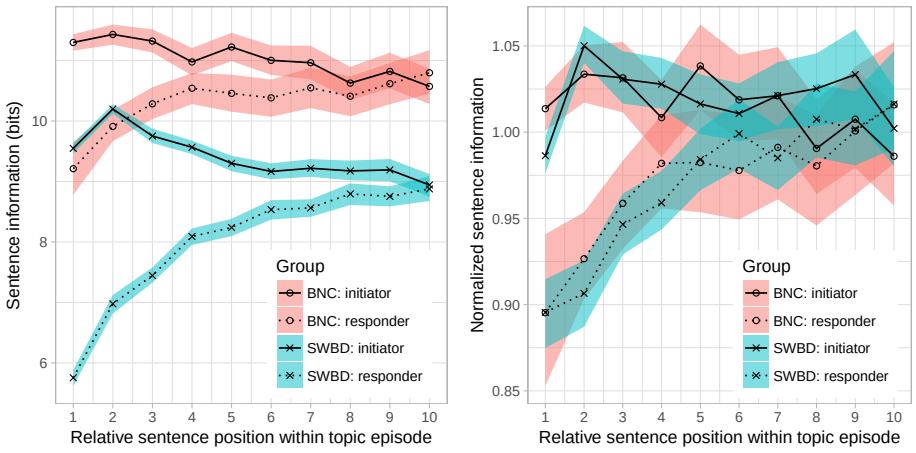
Therefore, we modify the operational definition of TIU for between-turn boundaries as follows: as many sentences immediately following the boundary as is necessary to reach a length threshold for the TIU, N. We set $N = 5$, the median of all first sentences after topic episode boundaries (regardless of turns). We explain the operational definition of TIU in Figure 3.

5.2. Results

We will discuss lexical information convergence before relating it to convergence of information contained in structural features of language.

5.2.1. Sentence information converges between topic initiator and responder

We distinguish the two speakers' roles in each topic episode: let the author of TIU be the *initiator* of the current topic, and the other one be the *responder*. We plot the sentence information (and normalized information) against its within-episode position respectively, grouped by speaker roles (initiator vs. responder) in Figure 4.



(a) Sentence information (b) Normalized sentence information

Figure 4: Sentence information (a) and normalized sentence information (b) against the relative sentence position within topic episodes, grouped by speaker roles (*topic initiator* vs. *topic responder*). Bootstrapped 95% confidence bands.

It can be seen that at the beginning of a topic episode, the initiators have significantly higher sentence information than the responders. As the topic develops, we can see that the initiators' information decreases (see Figure 4a, except for the temporary short increase from within-topic position 1 to 2 in SWBD) or stays relatively steady (see Figure 4b); and the responder's information increases. Together they form a convergence trend within topic episode.

We use linear mixed models to examine the observed convergence trend, i.e., to test whether the sentence information reliably decreases for initiators and reliably increases for responders. Models are fitted for initiators and responders respectively, using the sentence information (and normalized information) as response variables, and the within-episode position as predictor (fixed effect), with a random intercept grouped by the unique episode index. Our models show that for the sentence information, the fixed effect of within-episode position is reliably negative for initiators (SWBD, $\beta = -3.6 \times 10^{-2}$, $p < 0.001$; BNC, $\beta = -2.9 \times 10^{-2}$, $p < 0.05$) and reliably positive for responders (SWBD, $\beta = 3.3 \times 10^{-1}$, $p < 0.001$; BNC, $\beta = 1.4 \times 10^{-1}$, $p < 0.001$). For the normalized sentence information, the fixed effect of within-episode position is insignificant for initiators, which means there is neither increase nor decrease, and is reliably positive for responders (SWBD, $\beta = 1.4 \times 10^{-2}$, $p < 0.001$; BNC, $\beta = 1.2 \times 10^{-2}$, $p < 0.001$). Thus, the convergence trend is statistically reliable.

The observation that topic *initiators*' sentence information decreases within topic episode is at first glance inconsistent with previous findings that assert an increase of sentence information in written text, which will be discussed in Section 7.

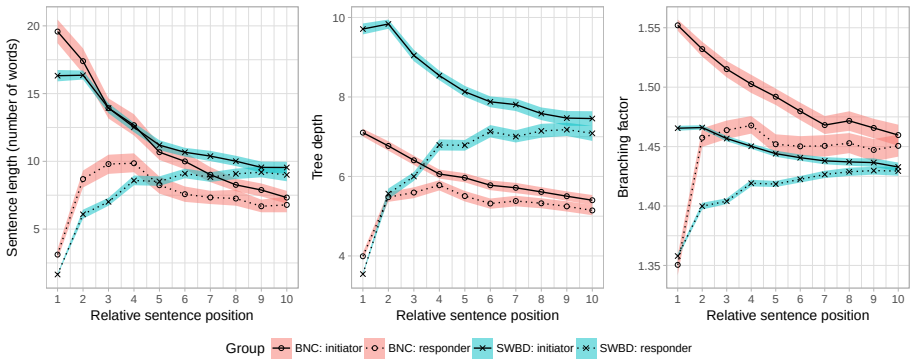


Figure 5: *Sentence length* (SL), *tree depth* (TD) and *branching factor* (BF) against within-topic sentence position (the relative position of a sentence from the beginning of the topic episode), grouped by speaker role, *initiator* vs. *responder*. Bootstrapped 95% confidence bands.

5.2.2. Convergence of other sentence-level measures

Previous studies have pointed out that sentence information is closely correlated with other syntactic complexity measures of sentence (Genzel & Charniak, 2002, 2003).

To examine whether the convergence of sentence information is accompanied by similar converging patterns of other measures of sentence-level linguistic representations, we consider three statistics: *sentence length* (SL), *tree depth* (TD), and *branching factor* (BF). SL is the number of words in a sentence. TD is the depth of the parse tree of a sentence. BF is defined as the average number of children nodes of all non-leaf nodes in the parse tree of a sentence. TD and BF are known to be positively and negatively correlated with sentence information in written text (Genzel & Charniak, 2003).

We plot the SL, TD, and BF of sentence against its within-episode position, grouped by speaker roles (*initiator* vs. *responder*) in Figure 5. All three measures show similar convergence patterns as sentence information (see Xu & Reitter 2016a for details). These results are expected, because the length and complexity of sentence are closely correlated with the information content. It also reflects the fact that the convergence of linguistic features within local topic episodes is ubiquitous in dialogue.

The convergence trend of entropy might be a reflection of alignment at lower representational levels. The interactive alignment model (IAM, Pickering & Garrod 2004) asserts that repeating words and syntactic choices between speakers will lead to increased alignment at higher linguistic representation levels, and we believe these results are compatible with that view. In the next section, we explore further from this perspective.

6. Linguistic Alignment as a Source of Information Convergence?

In understanding the mechanisms that may lead to speaker contributions with converging sentence information, the IAM framework (Pickering & Garrod, 2004) provides a possible explanation: the convergence of sentence information could be due to a process in which interlocutors gradually adopt each other's language at multiple levels (lexical, syntactic etc.). This process may, and that would be new, be constrained by topic episodes. Based on the well-known effects of syntactic adaptation (Bock, 1986), syntactic alignment has been related to increased task success in task-oriented dialogue (Reitter & Moore, 2014). Given that information needs to be effectively transmitted for such task success, we ask whether informational convergence co-occurs with some forms of alignment.

To validate this explanation, we need to examine whether the degree of linguistic alignment changes within topic episode, in a direction that is consistent with sentence information. In other words, if we model short-term alignment as a convergence process that, to some extent, restarts with each topic segment, we would expect increasing overlap towards the end of each segment if local alignment parallels the information pattern.

Due to space limitations, we only discuss alignment at the lexical level here. The *local linguistic alignment* (LLA) proposed by Fusaroli et al. (2012) and Wang et al. (2014) is used to quantify the strength of alignment. LLA is computationally defined as the number of repeated words between two bodies of text, *prime* and *target* (as P and T in Equations 4 and 5), normalized by their length:

$$\text{LLA}(P,T) = \frac{\sum_{w_i \in P} \delta(w_i, T)}{length(P) * length(T)} \tag{4}$$

$$\delta(w_i, P) = \begin{cases} 1, \text{ if } w_i \in P \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

We examine whether LLA increases within topic episode. If we observe a reliable increasing trend of LLA, then that would support our hypothesized explanation that similar linguistic representations shared between interlocutors contribute to the converged entropy. A linear mixed model was fitted using the relative position of $\langle P,\ T\rangle$ pair within topic episode as predictor, with a random intercept grouped by unique topic episode. Indeed, we found a significant effect of the relative position of utterance pairs, and the coefficient is positive (SWBD, $\beta = 2.2 \times 10^{-4},\ p < .001$; BNC, $\beta = 4.6 \times 10^{-4},\ p < .001$), which suggests that the lexical alignment between interlocutors does increase within topic episode.

Considering the fact that LLA is sensitive to the size of P and T, i.e., utterance length (Doyle et al., 2016), and that sentence length systematically changes in topic episode (see Figure 5), we come up with a normalized variant of LLA (in a similar way as to normalized sentence information), nLLA, which is independent of utterance length (see Appendix D for details). Same significant effect of the utterance pair position is found when nLLA is used to fit the linear mixed-effect models (SWBD, $\beta = 9.9 \times 10^{-3},\ p < .001$; BNC,

Figure 6: Normalized Local Linguistic Alignment (nLLA) against utterance position within topic episodes. Bootstrapped 95% confidence bands.

$\beta = 7.1 \times 10^{-3},\ p < .001$). We plot nLLA against the relative utterance position in Figure 6. Except for the case of Episode 1 in SWBD, we can observe the slight, yet consistent increase of nLLA within topic episodes.

These results supports the explanation that the convergence of sentence information between interlocutors may be facilitated by the increased alignment of linguistic representations (in this case, words). Of course, we show a correlation, but a causal analysis will require either a temporal-causality framework or an experimental design.

7. Discussion

7.1. Summary

The motivation of this study is rooted in the growing body of work on the principle of entropy rate constancy, and in the question of whether this principle generalizes to dialogue. We examine the variation patterns of sentence information in dialogues and how these patterns interact with the topic structures and speaker roles. This allows us to obtain a more focused picture of people's communicative behavior from an information-theoretical perspective.

What we found are new patterns of sentence information that are quite different from those in written text. Specifically, when distinguishing the speakers' roles by topic initiator vs. responder, the initiator's information decreases whilst that of the responder increases within each topic episode, and together they form a convergence pattern. The downward trend among topic initiators seems to be contrary to the ERC principle, but as we will discuss next, it is actually an effect of the unique topic structure schema of dialogue.

From an information-theoretic perspective, our findings provide an angle to view dialogue as a process of information exchange. To use classic language by Shannon (1948), in dialogue, interlocutors play the roles of information provider and receiver interactively within each topic episode. From the perspective of linguistic behavior, our findings provide further support to the existing work on convergence of linguistic representations in conversations. Both of the two perspectives will be discussed.

Beyond the effect of speaker roles, we do observe that sentence information increases with its global position in the dialogue, which is consistent with written text data(Genzel & Charniak, 2002, 2003; Qian & Jaeger, 2011; Keller, 2004). This indicates that human communication in spoken form does follow the general principle of constancy rate.

The distribution of information density among the two types of dialogic contributions may be a consequence of the cognitive load imposed by topic-shifting. Under an ERC/UID assumption at the level of the individual, one would need to postulate that topic initiation requires fewer cognitive resources for the expert topic-introducer, who retrieves the most salient, best-known facts to convey first; hence more information density is warranted. However, at the surface level, information density is not constant for the topic initiator's language within a topic episode. Only if we treat the dialogue partners as a system does information density follow the patterns proposed by the ERC/UID hypothesis.

7.2. Dialogue as a process of information exchange

By combining topic segmentation techniques and fine-grained discourse analysis, one may view human communication from a new vantage point: because interactants disseminate information in dialogue, we need to focus on the distribution of information density in order to describe strategies of speakers and the resulting structure in dialogue.

One difference between written and spoken language in conversation is that there is only one direct input source of information in the former, i.e., the author of the text, but for the latter, there are multiple and changing direct input sources, i.e., the different speakers. Because of this inherent difference, when language production is treated as a process of choosing proper words (or other representations) within a context, the definition of "context" is different between the two forms of communication. In written language (see Equation 4 in Section 2), C_i, the global context of a word X_i, is assumed to be all the words in preceding sentences. This is a reasonable hypothesis within UID, because for example, when an author writes a book, he will organize the distribution of information in sentences and paragraphs evenly for the benefit of the readers, i.e. the potential comprehenders, so that the later chapters can be inferred from previous ones, and the later paragraphs in a chapter can be inferred from earlier ones and so on. The key behind this delicate design in written text (if we accept UID as by design) is that constant contextual entropy maximizes the chance of successful comprehension (Jaeger, 2010).

In the case of dialogue, however, interlocutors switch rapidly between the roles of language producer and comprehender, and thus the concept of context is different. Within a dialogue, for any upcoming utterance, all preceding utterances together can be viewed as the *shared* context for the two speakers. We propose a mental experiment to help illustrate how the shared nature of context affects the interlocutors' behavior: Suppose we, as researchers and "super-readers", observe the transcript text of a dialogue between speaker *Alice* and *Bob*. To us, any upcoming utterance is based on the context created by all previous utterances, which is why we can observe a consistent increase

of sentence information within the whole dialogue. Also, to us, a new topic episode in dialogue is just like a new paragraph in written text, within which we can observe steady information increase without differentiating the utterances from the two speakers. But from the perspective of an individual interlocutor, he or she will not necessarily leverage the preceding utterances as a coherent context. Specifically, Alice, as a topic initiator, might not rely much on the previous shared context (say, the previous topic episode) as she introduces the new information that is from an "outer" context. Therefore her sentence information starts high (i.e., contains more information) and gradually decreases, because the new content she introduces (from an outer context) has low mutual information with the current context represented by the preceding utterances. On the other side, Bob as a topic responder relies on the previous shared context, including the initiator's very first utterance that introduces the new topic (i.e., TIU). Right after the TIU, he notices that his partner wants to talk about something else, which means the context of his next utterance has changed. It explains why his sentence information starts low – the mutual information between the upcoming utterance and the old context is reduced. Moreover, as the responder acts more as a passive follower at the early stage of the topic shift, he tends to take whatever is provided by the initiator and uses this to incrementally update the context. As the context is built with the new topic being developed, the mutual information between his next utterance and the context also increases, which causes the increase of sentence information.

Another angle to view the difference between dialogue and written text is that the former is dynamically constructed, where the participants have the opportunity to fix misunderstandings in real time, while the latter needs to be well designed in advance for the optimal chances of comprehension. Therefore, Alice, the topic initiator, can provide much information without worrying if Bob can comprehend it easily for she knows that he can always ask for clarification. Rather, it is Bob's reaction that matters to the quality of communication: if he also produces a lot of information immediately after Alice, then it will cause comprehension difficulty for both speakers.

To sum up, dialogue can be viewed as a communication system of two components that interactively exchange information between each other. The channel between the two components are efficiently used in a way that maximizes use of its capacity: when the information transmission rate from one side to the other is high (i.e., when the initiator is introducing a new topic), the rate of the other direction is correspondingly low (i.e., the responder is contributing less information early on in a topic episode); as the transmission rate of one direction decreases, the rate of the other direction increases, while the total absolute rate keeps constant. We believe that there is a cognitive mechanism behind this automatic adjustment of relative channel size. The respective cognitive load imposed by following the dialogue in a new direction may be complemented by reduced information in the form of language. This is, again, compatible with a communication framework that imposes a tendency to limit or keep constant overall information levels, but views the dialogue partners as a socio-cognitive system.

7.3. Grounding driven by the need for efficient information exchange

Effective communication calls for the success in building common ground between people. The converging patterns of sentence information found in this study reflects the process of grounding. If we break down the process of how common ground is built during the early phase of dialogue (or, segments in dialogue), we can find that the grounding process is driven by the need for efficiency in the joint work that is dialogue.

In the dialogues we have examined, when a new topic starts, the initiator knows best about this topic; the initiator contributes more information in her turn, hence its higher entropy. The responder at first knows little about the new topic; the purpose of his early utterances, from the perspective of grounding, is to let the initiator know that he has received the information of the new topic, which is why these utterances are simpler and more common ones that contain less lexical information, such as short acknowledging back-channel utterances, and short comments or queries and so on. As the conversation evolves, mutual knowledge, i.e., common ground, is accumulated, which means the responder knows more about the new topic (and he is certain that the initiator knows this, too), now in his feedback he can express more substantial opinions that contains more lexical information, i.e., higher entropy. On the other hand, the decrease of initiator's sentence information can be explained by the "drying up" of information – it is difficult and unnecessary for the initiator to keep maintaining high novelty in her contribution to the conversation because she also needs to acknowledge the responder's utterances; her own cognitive load also places limits on the duration of the act of "bringing in new content".

The result is a rational outcome for the system of interlocutors: it optimizes truth maintenance among dialogue participants, and it would predict more successful joint work when a dynamic information density is complementary among speakers. Suppose the rules of engagement would not allow for changing information density in individual speakers. Let's imagine, Alice and Bob split their use of the communication channel "fairly" in that they aim to contribute information at the same rate each. Alice is the expert, and Bob is a novice. Then, the outcome is unlikely to be the most truthful reflection of the dialogue's topic, nor does Alice transfer some of her knowledge to Bob at the optimal rate. (As an aside, examples of this can be found in political discourse or media representations of scientific debates, where minority dissenters or uninformed candidates are given similar airtime, with results that do not reflect the state of the art, or result in public opinion of a candidate that might not reflect a more neutral assessment.)

By distinguishing *topic initiators* from *topic responders* our model also reflects some of the key characteristics of joint activities. For example, as Clark (1996) proposed: in a joint activity, participants play different roles; an activity is usually comprised of sequences of sub-activities, and the participants' role may differ from sub-activity to next. Third, to achieve the goal of the activity, coordination between participants of different roles is required. In our case, a *topic episode* within a dialogue can be viewed as a sub-activity, in which the

initiator sets up a dominant goal, i.e., to develop a new *topic*, and the responder joins him in order to achieve the goal. The role of an initiator is indicated by her high activity (more production and high entropy) at the early stage, while the role of a responder requires him to take a more "accepting" stance. The roles of initiators and responders are not fixed among interlocutors across topic episodes in a dialogue. The initiator of topic could be the responder of next, and vice versa.

7.4. Linguistic alignment parallels the convergence of sentence information

The convergence of sentence information within topic episodes can be interpreted as a sign of interactive alignment between the interlocutors. The Interactive Alignment Model (IAM, Pickering & Garrod 2004) predicts that speakers tend to adopt their interlocutor's choices of words and syntactic rules. Increased repetition of linguistic features directly increases the similarity of language productions between interlocutors. Thus the convergence can be a consequence of increased lexical and syntactic repetition.

Note that alignment is stronger for less expected, lower-frequency linguistic elements (Jaeger & Snider, 2013; Reitter et al., 2011), so it is not surprising to assume that speakers attend more to novel, information-rich material and use this material as a source of adaptation. Note, though, that the calculation of sentence information in our study relies on the lexical context rather than a notion of *surprisal* that implies more complex operations that build an expectation about future words. So, the model so far does not assume error-driven learning in calculating what constitutes information, but rather fundamental cognitive mechanisms of memory retrieval (c.f., Kaan & Chun 2017).

Nevertheless, our findings reflect that lexical linguistic alignment (or, accommodation, coordination) is not just observed within the scope of the whole dialogue, but may also happen within shorter units, i.e., topic episodes (c.f., Reitter & Moore, 2014). The relatively large discrepancy of sentence information at the beginning of topic episodes[2] suggests that the two interlocutors diverge at first in their linguistic choices and later become aligned (coordinated) towards each other as the topic develops. It seems that when initiating a new topic, interlocutors can no longer rely on the previously achieved aligned local lexicon. Linguistic style at all representational levels (lexical, syntactic, information density, semantic and syntactic complexity, and so on) may be reset or at least reduced. This is a prediction inferred from the observation of converging sentence information, which suggests new experiments to test such a hypothesis of the scope of alignment.

The convergence of sentence information between the interlocutors in two roles suggests an automatic process by which the interlocutors coordinate the information content of their production with what they have perceived from their partners. This coordination at information content level can be related

[2]Even discounting the low information density in topic episode 1 in Switchboard, which may reflect conventional greetings.

to alignment that is found at other higher levels of linguistic representations, such as speech rates (Webb, 1969), or the intent of the speech acts (Wang et al., 2015). More abstract yet, the fragmented topic episodes in dialogues can be seen as the locus where interlocutors build temporarily shared understanding (Linell). This has even been characterized as a "synchronization of two streams of consciousness" (Schutz, 1967).

These convergence phenomena are hardly independent of each other (we show lexical alignment within topic episodes in this paper). For example, consider the duration of dialogue turns. Sentences of higher information tend to be longer. The converging pattern indicates that the gap between the two interlocutor's turn occupancies narrows over time. A "speaker" becomes more of a "listener", and vice versa. We think that even if some of these convergence processes are ephiphenomena of others, they serve a purpose in creating mutual understanding. For example, Fusaroli et al. (2014) and Reitter & Moore (2007, 2014) show, in different languages and using different methods, that lexical and syntactic adaptation are related to task success. In separate, recent work, we examine higher-order overlap of the frequencies with which interlocutors shift between information-high and information-low states (Xu & Reitter, 2017). By transforming the metric into frequency space, we assume that several parallel cognitive processes contribute to the ebbs and flows of information. When we take the phase shift between interlocutors into account, we can characterize how well information contributions complement each other in all of these frequencies. Overlap in frequency space and complementary in phase can then predict how successful participants are in their task, in task-oriented dialogue. The periodic nature of information convergence (along topic boundaries) could be seen as another link in the chain of observations of convergence at different representational levels.

8. Conclusion

This study was motivated by a desire to describe dialogue as a process of information exchange that can be quantified, and that is managed by interlocutors using alignment and/or grounding. The two models of achieving mutual understanding were not contrasted here. Instead, we used both of them to explain the data, assuming that they both apply to dialogue at different levels and in different circumstances. One conclusion is that all three descriptive models are subjugated to the topic structure of dialogue – a structural configuration that may involve many more layers than shown in our study. Is it rules for information distribution that cause speakers to align and ground, or is the information density distribution the result of alignment and grounding? Neither of those options needs to be the case. The different models of dialogue are empirically related, as we show in this paper, and we argue for basic cognitive bounds and mechanisms of general learning and memory retrieval as an explanation of observable distance and convergence in information density, alignment and the conventions of grounding.

To reach these conclusions we measured the information density in dialogue using the averaged per-word information within sentence, which then led us to characterize the turn-taking in the exchange of information between the conversation partners. We validated the principle of entropy rate constancy in spoken dialogue, using two common corpora. Besides the results that are consistent with previous findings on written text, we find new patterns unique to dialogue. Interlocutors who actively initiate a new topic tend to use language with higher information, compared to the language of those who passively respond to the topic shift, which together shapes the convergence of information density as the topic develops. We explain this observed convergence of information density from the perspective of information exchange, the process of grounding, and the alignment of linguistic behaviors.

By showing that ERC applies to the system level rather than to the individual speaker in dialogue, we work towards a unified perspective where the low-level linguistic alignment behavior and the high-level joint activity nature of dialogue are combined. This perspective may eventually lead us to identify further systematic patterns of interaction in verbal communication.

Acknowledgments

The National Science Foundation supported this research (IIS-1459300, BCS-1457992). We thank Frank E. Ritter and several reviewers for comments on earlier versions of this manuscript and of Xu & Reitter (2016b), which presented preliminary work leading to this study.

References

Barr, D. J., & Keysar, B. (2004). Is language processing different in dialogue? *Behavioral and Brain Sciences*, *27*, 190–191.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, *67*, 1–48.

Blei, D. M., & Moreno, P. J. (2001). Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 343–348). New Orleans, LA.

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

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition*, *75*, B13–B25.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, *13*, 127–149.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Koshi, B. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.

Doyle, G., & Frank, M. (2015). Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1587–1596). Denver, CO.

Doyle, G., Yurovsky, D., & Frank, M. C. (2016). A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 637–648). Montreal, Canada.

Eisenstein, J., & Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 334–343). Honolulu, HI.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological Science*, *23*, 931–939.

Fusaroli, R., Raczaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 199–206). Philadelphia, PA.

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 65–72). Sapporo, Japan.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing* (pp. 517–520). volume 1.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Stroudsburg, PA.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, *23*, 33–64.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Jaeger, T. F., & Levy, R. P. (2006). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems* (pp. 849–856).

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*, 57–83.

Johnson, R. A., Wichern, D. W. et al. (2014). *Applied multivariate statistical analysis* volume 4. Prentice-Hall New Jersey.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling, . `arXiv:1602.02410`.

Kaan, E., & Chun, E. (2017). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition*, (pp. 1–15).

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*, 400–401.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 317–324). Barcelona, Spain.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Linell, P. (). *Approaching dialogue: Talk, interaction and contexts in dialogical perspectives*. IMPACT: Studies in Language and Society.

Linell, P. (1990). The power of dialogue dynamics. In I. Markovà, & K. Foppa (Eds.), *The dynamics of dialogue* chapter 8. (pp. 147–177). Hertfordshire, England: Harvester Wheatsheaf.

Malioutov, I., & Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 25–32).

McCarthy, J. (1987). Formalization of two puzzles involving knowledge. In V. Lifschitz (Ed.), *Formalizing common sense: Papers by McCarthy* (pp. 158–166). Norwood, NJ: Ablex Publishing.

Ng, S. H., & Bradac, J. J. (1993). *Power in language: Verbal communication and social influence*. Sage Publications, Inc.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*, 2382–2393.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

Qian, T., & Jaeger, T. F. (2011). Topic shift in efficient discourse production. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3313–3318). Boston, MA.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*, 587–637.

Reitter, D., & Moore, J. D. (2007). Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 808–815). Prague, Czech Republic.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46.

Schiffer, S. R. (1972). *Meaning*. Oxford University Press.

Schutz, A. (1967). *The phenomenology of the social world*. Northwestern University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics 9: Pragmatics* (pp. 315–332). New York: Academic Press.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *The 7th International Conference on Spoken Language Processing, ICSLP-INTERSPEECH*. Denver, CO.

Temperley, D., & Gildea, D. (2015). Information density and syntactic repetition. *Cognitive Science*, *39*, 1802–1823.

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

Vega, A., & Ward, N. (2009). Looking for entropy rate constancy in spoken dialog. In *Departmental Technical Reports (CS)*. Paper 46.

Wang, Y., Reitter, D., & Yen, J. (2014). Linguistic adaptation in online conversation threads: analyzing alignment in online health communities. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics (CMCL). Workshop at the Meeting of the Association for Computational Linguistics (ACL 2014)* (pp. 55–62). Baltimore, MD.

Wang, Y., Yen, J., & Reitter, D. (2015). Pragmatic alignment on social support type in health forum conversations. In *Proceedings of Cognitive Modeling and Computational Linguistics (CMCL). Workshop at the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)* (pp. 9–18). Denver, CO.

Webb, J. T. (1969). Subject speech rates as a function of interviewer behaviour. *Language and Speech*, *12*, 54–67.

Xu, Y., & Reitter, D. (2016a). Convergence of Syntactic Complexity in Conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 443–448). Berlin, Germany.

Xu, Y., & Reitter, D. (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* (pp. 537–546). Berlin, Germany.

Xu, Y., & Reitter, D. (2017). Spectral Analysis of Information Density in Dialogue Predicts Collaborative Task Performance. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada.

Appendix A. Selection of Training Set for the Language Models

We compare three different ways of training the LMs that will be later used to compute sentence information. We use two commonly used indices, perplexity (*ppl*) and the number of out-of-vocabulary words (OOVs) to evaluate the performance of the LMs.

Appendix A.1. Training from external corpora

We train an LM from all the sentences in SWBD, LM_{SWBD}, and use it to compute the sentence information in BNC. Similarly, we train an LM from BNC, LM_{BNC}, and use it on SWBD. It turns out that the performance of LMs is not good: When tested on BNC, the *ppl* of LM_{SWBD} is 266.2, with 74.1 K OOVs. When tested on SWBD, the *ppl* of LM_{BNC} is 179.8, with 141.7 K OOVs. These *ppl* values are pretty big, compared to the performance of the recently reported state-of-the-art language models (Jozefowicz et al., 2016), and the OOVs also take up nearly 10% of the words in testing set.

We also trained an LM from a larger corpus, the thread conversations in the Cancer Survivor Network (CSN), which contains about 40 M words. When we test this LM_{CSN} on SWBD, the *ppl* is 244.4 and the number of OOVs is 108.4 K. When tested on BNC, the *ppl* is 266.1, but there are only 7.1 K OOVs.

Appendix A.2. Non-position-wise cross-validation

We use the 10-fold cross-validation method to train LMs, but do not differentiate sentence positions. I.e., after randomly divide the corpus into 10 subsets, S_i $(i = 1, 2, \ldots, 10)$, for each round of cross-validation, we just train one LM from $\{S_j | j \neq i\}$, and use it to compute the information of all the sentences in S_i (disregarding their positions).

Within this process, we trained 10 LMs on each corpus. For SWBD, the LMs' average *ppl* is 77.4 ($SD = 1.9$), and total number of OOVs is 13.0 K. For BNC, the average *ppl* is 107.4 ($SD = 15.5$), and the total number of OOVs is 14.4 K.

Appendix A.3. Position-wise cross-validation

We then carry out the position-wise cross-validation that is described in Section 3.1. Now we have 100 LMs trained in each round of cross-validation, and the 10-fold setup results in 1000 LMs for each corpus. For SWBD, the LMs' average *ppl* is 89.0 ($SD = 9.6$), and number of OOVs is 115.0 K. For BNC, the average *ppl* is 84.9 ($SD = 15.6$), and number of OOVs is 55.4 K.

In summary of the results, the LMs trained from external corpora, i.e., LM_{SWBD}, LM_{BNC} and LM_{CSN}, have the poorest performance in terms of *ppl* and OOVs. The LMs trained by the non-position-wise cross-validation have the lowest *ppl*s and OOVs numbers. However, to be consistent with the previous work in methodology, and considering that the *ppl*s of the LMs are acceptable, we decide to use the position-wise cross-validation method in this study.

Appendix B. Alternative Topic Segmentation Algorithms

Appendix B.1. Descriptive Statistics

First of all, we compare the statistics (mean, median, and standard deviation) of the length of topic episodes (number of sentences) resulted from the three algorithms (table B.4). It shows that the segments generated by `BayesianSeg` and `MinCutSeg` are longer in length, but with larger *SD*. `TextTiling` generates shorter segments with smaller *SD*. We also notice that `BayesianSeg` and `MinCutSeg` generate much smaller median values (4.0 and 1.0) in BNC, which potentially undermines the validity of these two algorithms because it is infeasible to consider a single sentence as a "topic". On the other hand, `TextTiling` gives smaller mean segment length for both corpora, and a reasonable median value (10) for BNC.

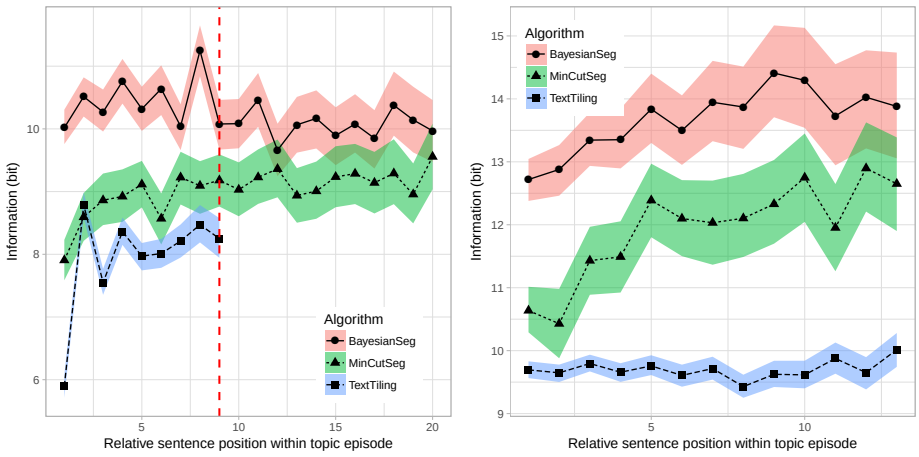
Table B.4: Basic statistics of the segment length (number of sentences) resulted from the three segmentation algorithms.

Algorithm	SWBD			BNC		
	Mean	*Median*	*SD*	*Mean*	*Median*	*SD*
`BayesianSeg`	19.7	9.0	27.0	13.1	4.0	35.1
`MinCutSeg`	19.7	7.0	26.7	13.1	1.0	29.9
`TextTiling`	9.2	8.0	5.7	10.9	10.0	7.3

Appendix B.2. Information patterns within topic episodes and near topic boundaries

One indicator of a topic shift in written text is that sentence information drops at the topic boundaries, e.g., the beginning of a paragraph (Genzel & Charniak, 2003) etc, and before encountering the next boundary, sentence information should keep increasing. In fig. B.7 we demonstrate how sentence information changes within the topic episodes resulted from the three algorithms respectively. To avoid overlapping the curves, we add 1 to the information values of the `BayesianSeg` curve, and minus 1 from the `TextTiling` curve. In SWBD, the mean segment length of `TextTiling` is shorter than the other two, which is reflected by the vertical cut-off line in fig. B.7a.

We use linear mixed-effect models to examine whether sentence information reliably increases with its relative position within topic episodes (random intercept grouped by distinct topic episode): `ent` $\sim$ `inTopicPos + (1|uniqueTopicId)`. The models show that for the segmentation results of `MinCutSeg`, the effect of the within-episode position on sentence information is significant: SWBD, $\beta = 5.1 \times 10^{-2},\ p < .001$; BNC, $\beta = 1.8 \times 10^{-1},\ p < .001$. The models on `TextTiling` also show significant effect, and the results are already reported in the original manuscript. For the segmentation results of `BayesianSeg`, the effect is only significant in BNC ($\beta = 1.2 \times 10^{-1},\ p < .001$), but not in SWBD ($p > .05$). However, since we do observe a pretty clear increasing trend within



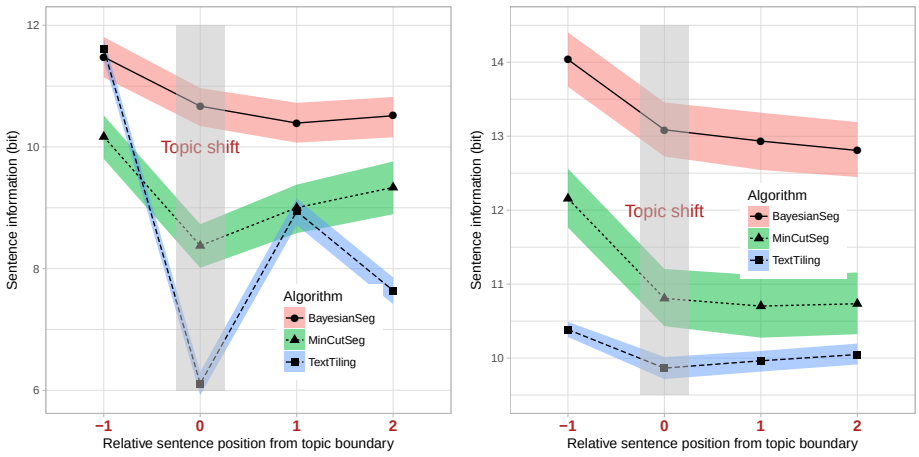
(a) SWBD (b) BNC

Figure B.7: Increase of sentence information within topic episodes. The red vertical dashed line in (a) denotes the 9th sentence. Bootstrapped 95% confidence bands.

the first 9 or so sentences (see the solid curve of fig. B.7a), we fit the model again just using that proportion of data (`inTopicPos` < 10), which gives a significant effect: $\beta = 5.5 \times 10^{-2}, p < .05$; if we use data in the range of $10 \leq$ `inTopicPos` ≤ 20, the effect is gone. Thus, it can be said that within the topic episodes resulted from applying `BaysianSeg` on SWBD, sentence information increases at first, and then drops a bit and eventually becomes stable.

Now that we can confirm that the increase of sentence information is captured by `BayesianSeg` and `MinCutSeg`, the next step is to examine how it changes near the topic episodes, hoping to observe a significant decrease of information from the end of the preceding episode towards the start of the next one. The decrease of sentence information from $x = -1$ (one sentence before topic shift) to $x = 0$ (where topic shift happens) can be observed, and further t-tests confirm the significance of the decrease: `BayesianSeg` on SWBD, $t(9685) = 3.51$, $p < .001$; `MinCutSeg` on SWBD, $t(8124) = 6.73$, $p < .001$; `BayesianSeg` on BNC, $t(10665) = 3.49$, $p < .001$; `MinCutSeg` on BNC, $t(10378) = 4.80$, $p < .001$.

To sum up the above results, `BayesianSeg` and `MinCutSeg`, as the alternatives of `TextTiling`, can effectively capture the characteristics of sentence information resulted from the *real* topic shift, i.e., sentence information drops when topic shift occurs, and gradually increases within topic episode. Therefore, using segmentation algorithms to identify the topic shift in conversations is a valid operation. Considering the fact that `BayesianSeg` utilizes word surprisal (negative log probability) as an input feature of the model (Eisenstein & Barzilay, 2008), which brings confounding effect to our observation of sentence information, and that `MinCutSeg` generates extremely small median value of segment length in BNC, we decide to use `TextTiling` to present the results in the manuscript, but keep the results from the other two algorithms in the



(a) SWBD (b) BNC

Figure B.8: The change of sentence information across topic boundaries. The x axis is the sentence position relative to the topic boundary: $x = 0$ is where topic shift happens (emphasized by the grey rectangle region); $x = -1$ is one sentence before, and $x = 1$ is one after, etc.

appendix.

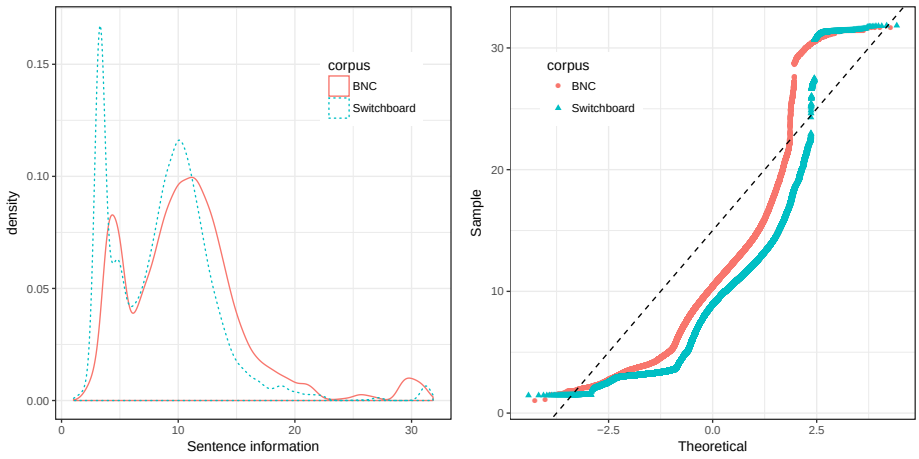
Appendix C. Distribution of Sentence Information

In Figure C.9 are the density curves and quantile-quantile (Q-Q) plots of the distribution of sentence information in and SWBD and BNC (First 100 sentences in each dialogue are selected). The density curves demonstrate two peaks and a slightly right-skewed shape. The Q-Q plots deviate much from the dashed straight line that indicates the shape of a normal distribution.

These plots suggest that the distribution of sentence information is not gaussian. A Shapiro-Wilk normality test (Shapiro & Wilk, 1965) shows that indeed the two distributions are significantly different from a normal (gaussian) one: SWBD, $W = 0.91,\ p < 0.001$; BNC, $W = 0.91,\ p < 0.001$.

Furthermore, we also examine the distribution of the normalized sentence information (See the definition of this normalized information in Section 3.2.2) in Figure C.11. The distribution of normalized information has even worse normality: the dual-peak remains in the density curves and the Q-Q plots show very large deviations. Shapiro-Wilk tests show that they are significantly different from a normal distribution: SWBD, $W = 0.60,\ p < 0.001$; BNC, $W = 0.82,\ p < 0.001$.

The possible reason that causes this larger deviation from normal distribution can be the way that the normalized information is computed: it is a transformation based on the length of sentence (number of words), while the



(a) Density curves (b) Q-Q plots

Figure C.9: Density curves (a) and quantile-quantile (b) plots of sentence information.

distribution of sentence length has a density curve that follows the quasi power law (See Figure C.10).

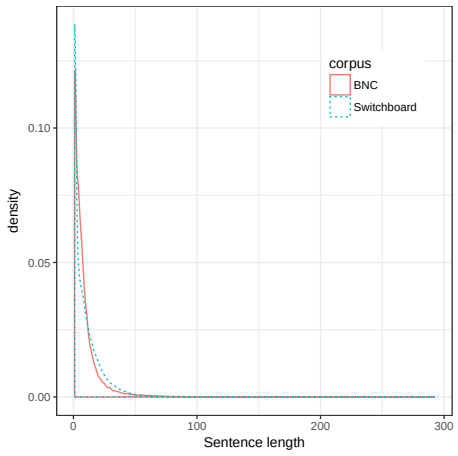
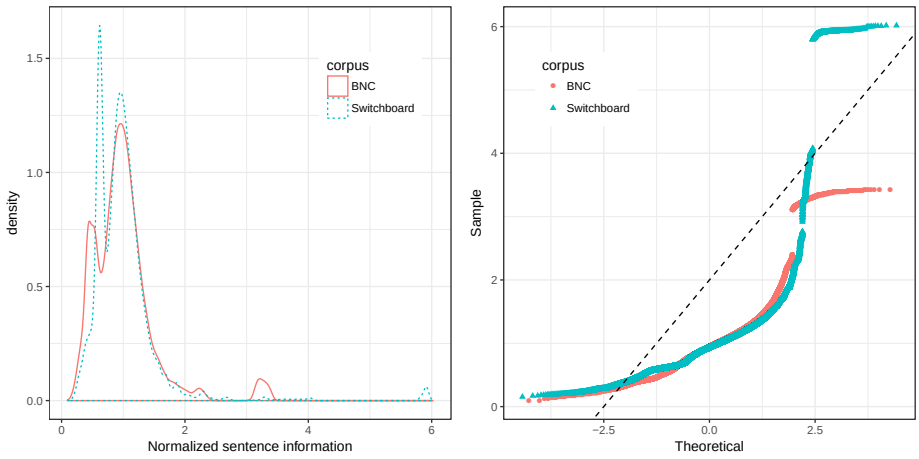


Figure C.10: Density curve of the distribution of sentence length (number of words in a sentence)

Next, we examine whether non-linear transformation can reduce the non-normality of the distribution. We obtain the logarithm (with base 2) of sentence information and normalized sentence information, and plot their density curves and Q-Q plots in Figure C.12 and Figure C.13 respectively. Although the density curves and Shapiro-Wilk tests indicate that the distributions of log-transformed sentence information and log-transformed normalized sentence information are



(a) Density curves (b) Q-Q plots

Figure C.11: Density curves (a) and quantile-quantile (b) plots of the normalized sentence information

still significantly different from a normal one, we can tell from the Q-Q plots that the logarithm transformation does improve the normality to some degrees.

Appendix D. Normalized Alignment Computation

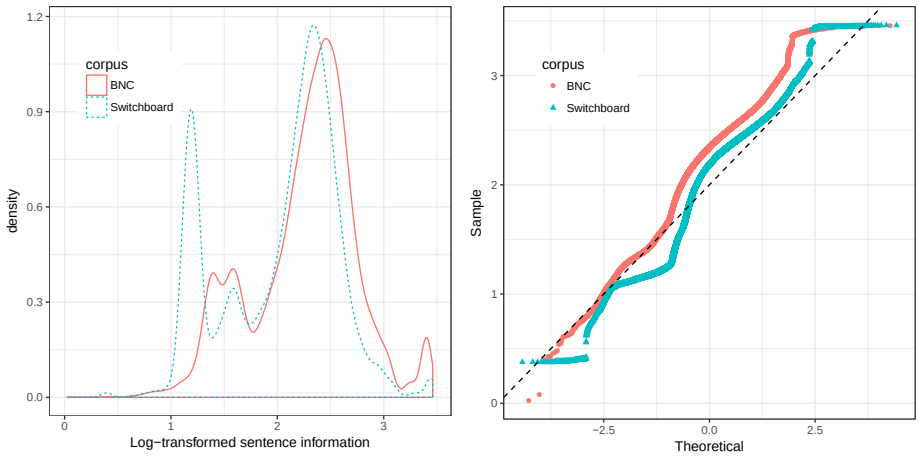
The *local linguistic alignment* (LLA) is sensitive to the length of utterances being calculated (Doyle et al., 2016). Longer P and T tend to result in smaller LLA, and vice versa. We define a measure of normalized LLA (nLLA) by eliminating the confounding effect of text length. We first compute the average LLA for all $\langle P,\ T\rangle$ pairs that have the same product of length (i.e., the denominator in Equation 4), $length(P) * length(T) = n$, and for all possible product values, $n = 1, 2, \ldots$:

$$\overline{LLA}(n) = \frac{1}{|S(n)|} \sum\nolimits_{\langle P_i,\ T_i\rangle \in S(n)} LLA(\langle P_i,\ T_i\rangle) \tag{D.1}$$

in which $S(n)$ denotes the set of $\langle P,\ T\rangle$ pairs that satisfy $length(P) * length(T) = n$. Then nLLA is computed by:

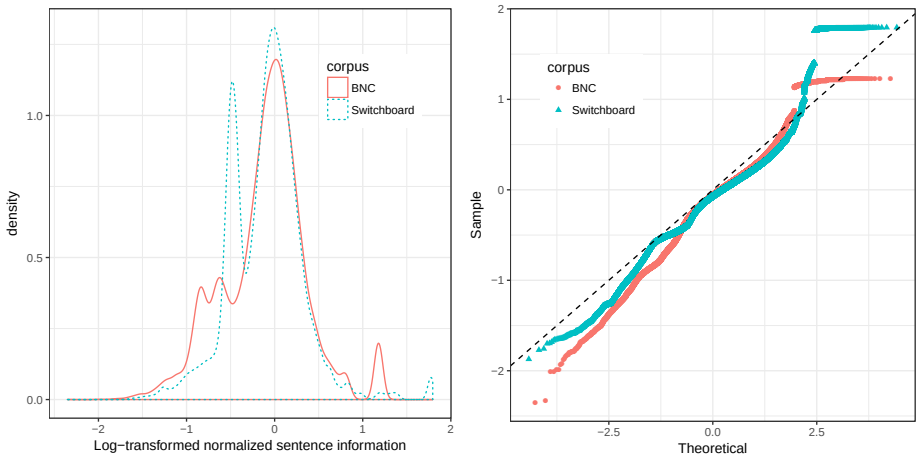
$$nLLA(\langle P,\ T\rangle) = \frac{LLA(\langle P,\ T\rangle)}{\overline{LLA}(length(P) * length(T))} \tag{D.2}$$

nLLA is not sensitive to length of P and T. The way we compute nLLA is adopted from Genzel & Charniak (2003), in which the normalized tree depth and branching factor of sentences were computed.



(a) Density curves (b) Q-Q plots

Figure C.12: Density curves (a) and quantile-quantile (b) plots of the logarithm of sentence information



(a) Density curves (b) Q-Q plots

Figure C.13: Density curves (a) and quantile-quantile (b) plots of the logarithm of normalized sentence information