

Theme-First Principle Revisited: Word Order, Information Structure, and Information Theory

Nobo Komagata

Department of Computer and Information Science

University of Pennsylvania

200 South 33rd Street

Philadelphia, PA 19104, USA

komagata@linc.cis.upenn.edu

Abstract

Although various forms of ‘theme-first’ principles have been proposed, its universality has been questioned with a number of counterexamples and the existence of arguably ‘rheme-first’ languages. Since theme-first principles still seems to play a significant role in accounting for word order in many languages, it is worth pursuing whether there is a principle that would be consistent with the observation. This paper applies an idea in information theory to the analysis of information structure and argues that word order with even distribution of informativeness is preferred. We will also see that counterexamples and ‘rheme-first’ languages are actually consistent with the proposed theory.

1 Introduction

The idea of theme-first principles is rather old, dating back at least to the eighteenth century [Lambrecht, 1994, p. 199]. Since then, there have been a number of proposals [e.g., Mathesius, 1975; Firbas, 1964; Halliday, 1967]. In the strongest form, such a principle would state that the sentence-initial position is reserved for a theme. Some proposals refer to notions such as ‘topic’ and ‘oldness’ instead of ‘theme’, and there surely are subtle differences between them. However, there is some common idea among them, and this paper stick to the term ‘theme’ as a cover term. A slightly weaker form would state that the theme always comes at the beginning of a sentence (weaker in a sense that a sentence does not necessarily have a theme). These proposals seem to be able to account for certain word order phenomena, especially in certain ‘free-order’ languages where syntax plays the lesser role.

Nevertheless, these theme-first proposals cannot be maintained in the forms stated above because there are a number of counterexamples. For example, Jespersen [1924] discusses the following example as reviewed in Lambrecht [1994, p. 50, Sec. 4.7]:

(1) *a.* Who said that?

b. **Peter** said it.

The sentence-initial position must be understood as the rheme of the response. Similar observations have been made by Steedman [2000, Sec. 3.1] as well.

Furthermore, Lambrecht [1994, p. 200] points out that a greater problem for theme-first principles is the existence of arguably rheme-first languages. For example, Mithun [1995] reports data from Siouan, Caddoan, and Iroquoian languages. We will now cite an example in Iroquoian (Tuscarora stories). The background is as follows. After the description of a long journey on the ice, discovery of land, and preparation for a sacrifice, the speaker introduces the head man, never mentioned before, at the beginning of the following

phrase (some phonetic symbols have been replaced for font availability reasons: ‘ǁ’ for right-hooked schwa and ‘ʔ’ for glottal stop).

(2) *haʔ uhǁʔnǁʔ ruʔnǁʔǁh, wahrǁhrǁʔ, ...*
the head man he said

“the headman said, ...”

Later, the speaker begins his recipe for cornbread. This time, a new (grammatical) object *ash* is introduced before the verb.

(3) *Tyhraetšihǁ kǁ:θ uhsǁéharǁh .. waʔkkúhaeʔ.*
first customarily ash I went after

“First, I usually would go after ashes.”

Similar data in other languages are also reported by Payne [1987] and Creider and Creider [1983]. Although it is not obvious that these are indeed rheme-first languages, the data still seem to show a consistent pattern rather different from *more* theme-first languages. Thus, the data deserve a closer examination.

Now that we cannot maintain theme-first principles, at least in the strong forms, could we still say something general about the relation between word order and information status? Counterexamples in languages like English do not seem to be abundant. In addition, the rheme-first languages seem to be limited to a small number of languages. How should we interpret this situation? If the different word order principles apply to different languages in an ad hoc way, it would pose a challenge to universal account of language as a human cognitive process. Since information structure (theme-rheme structure) has been associated with word order in various forms, esp. by the Prague school [e.g., Sgall et al., 1986], the above observation may undermine the role of information structure. Furthermore, Natural Language Processing systems might never be able to contain a base word-order module that could be the basis for all languages. One way to deal with the situation is to consider different notions of, say, ‘theme’s as in Kruijff-Korbyová et al. [2000, Sec. 1.4.2]. Lambrecht [1994, p. 202] too distinguishes accented and non-accented themes (‘topic’ in his term). Although such a move might eventually be necessary, it seems preferable to push the limit and explore the possibility of finding some general underlying principle that could be applied to a wide range of research proposals.

The starting point of this preliminary paper is Vallduví [1990, p. 15]. He cites Dretske [1999] about the use of the notion of ‘information’, but never develops the idea any further. I will take full advantage of information theory as done by Dretske, and will analyze word order from that point of view. In this connection, I will also discuss the definition of information structure from an information-theoretic point of view.

The main hypothesis discussed here is that there is a way to view the theme-first tendency without being suffered from apparent exceptions. Specifically, I will first consider information structure as a means to even out the information load carried by the theme and the rheme. In other words, we will interpret information structure as a means to minimize the standard deviation of the entropies of the theme and the rheme, which is called ‘information balance’ in this paper.

One of the consequences of this hypothesis is that it is always optimal if we deliver the less informative (i.e., low-entropy) component before the more informative (i.e., high-entropy) one. If we assume that information structure is a division of components so that the theme has lower entropy than the rheme, this can be the universal principle behind the theme-first tendency.

There are a few cases where the ordering does not matter. The most important, as it seems, is the following: if one component is totally predictable (i.e., zero entropy), the ordering does not affect the information balance. Presumably, the predictable component is the theme. I hypothesize that this situation corresponds to the apparent exceptions to the theme-first principle.

This paper is still in a preliminary stage. My intention is to solicit as many inputs as possible from

colleagues to refine or modify the idea. Since I will make a fairly straightforward connection between information structure and information theory, it may be easy to spot problems with the connection. But I hope that the current approach would at least present precisely-stated hypotheses for well-focused discussion.

One thing the current proposal has nothing to say is why the rheme-theme is used in certain cases (when the present proposal does not predict either ordering). Word order is a complex phenomenon. Each language imposes various lexical, syntactic, and pragmatic constraints on word order. It is not surprising that information structure does not explain all the word-order phenomena. In addition, we limit the discussion of constituent ordering, not word ordering within a phrase, where morpho-syntax tends to fix word order quite rigidly.

The rest of this note is organized as follows. Section 2 introduces the information-theoretic hypotheses as much informally as possible. The corresponding mathematical treatment is included in Appendix A. Section 3 discusses various rheme-first cases and analyzes whether they are accountable within the current approach. This section is rather premature and will be extended with more analyses. Section 4 presents an information-theoretic definition of information structure as a basis for the rest of the paper. Section 5 concludes with possible future work.

2 Information-Theoretic Account of Theme-First Tendency

In this section, we informally discuss the idea of applying information theory to the analysis of theme-first tendency. We will use some mathematical notations because certain properties are succinctly represented that way. But the actual mathematical treatment including a proof of the main theorem is given in Appendix A.

The main point here is that we can apply the idea of informativeness based on the measure of ‘entropy’. Note that informativeness is considered in different ways in different approaches. For example, in the formal semantics tradition, informativeness is a relation on a hierarchical structure (mathematically, a lattice) of semantic representations. The use of entropy in the present approach is different from using such a structure because information is measured as numeric values.

The use of entropy has been discussed in relation to linguistics and philosophy for a long time [Bar-Hillel, 1964; Crosson and Sayre, 1967; Cherry, 1978]. While some suggests usefulness [Cherry, 1978], some are more cautious [Bar-Hillel, 1964] saying that information is different from ‘meaning’. We do not attempt to capture the meaning of an utterance using entropy. What we do here is to demonstrate that entropy can be used to account for certain word order phenomena.

The basic idea about entropy is fairly simple. Roughly, if we have more options (possibilities), the entropy of the event is higher. Informally, a high entropy is associated with high informativeness, low predictability, high uncertainty, more surprise, etc. In the simplest scenario where each possibility of the event is equally likely, the entropy of the event is directly related to the number of choices. In terms of probability, the chance of hitting a particular choice out of n choices is $1/n$. Entropy is a measure related to this probability with a non-negative value, but it is adjusted (logarithmically) so that the effect of the increase in the number of choices is more in accordance to human sense. We could compare this with a measure of the loudness of sound. The effect of increasing the volume level of an audio system decreases even though we increase the volume proportionally.

For example, suppose that the rheme of an utterance is a predicate that must be chosen from 3 possible properties, say, *tea*, *coffee*, and *soda*. Also consider another case where the option involves 5 such possibilities: *large*, *wooden*, *flat*, *expensive*, and *purple*. If the distribution is even, the latter event has a higher entropy.

Entropy is a general function that can be applied to an arbitrary probability distribution, not necessarily even distributions. For an event X with an arbitrary probability distribution, let us denote its entropy as

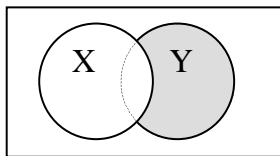


Figure 1: Conditional entropy

$H(X)$. If the probability distribution is uneven, the increased number of choices does not necessarily mean higher entropy. But for the sake of the present discussion, we can just imagine even distributions.

We now extend the use of entropy to two events, which will be applied to theme and rheme shortly. If the two events, X and Y , are completely independent, or have no effects on the other, the total entropy (called ‘joint entropy’) of the two events, written as $H(X, Y)$, is just the sum of the two entropies, i.e., $H(X, Y) = H(X) + H(Y)$. On the other hand, if the two events are completely dependent, the joint entropy and the entropy of the individual events are all the same, i.e., $H(X, Y) = H(X) = H(Y)$. Between these two extreme cases, there must be some dependence between the two events. Thus the joint entropy is somewhere between the sum of the entropies of the two events and the entropy of the more predictable event, i.e., $H(X) \leq H(X, Y) \leq H(X) + H(Y)$, if $H(X) \leq H(Y)$.

In general, we can describe the two-event situation as follows: the joint entropy is the sum of the entropy of one event and that of the other event depending on the first event, i.e., $H(X, Y) = H(X) + \langle \text{entropy of } Y \text{ depending on } X \rangle$. We call $\langle \text{entropy of } Y \text{ depending on } X \rangle$ ‘conditional entropy’ and denote it as $H(Y|X)$. Thus, we have $H(X, Y) = H(X) + H(Y|X)$. The relation between these measures is shown in Fig 1. The entire area covered by X and Y corresponds to $H(X, Y)$, the area for X corresponds to $H(X)$, and the shaded area of Y excluding the intersection of X and Y corresponds to $H(Y|X)$. If the events are analyzed in different order, we have $H(X, Y) = H(Y) + H(X|Y)$.

Based on the above basic ideas about information theory, we now apply them to the analysis of information structure and word order. Let us assume that an utterance is partitioned into a theme and a rheme as in the second utterance in the following example.

(4) *i.* John has a house.

ii. [The door]_{Theme} [is **purple**]_{Rheme}.

Suppose that immediately after the first utterance, the speaker wants to deliver some proposition. She might have thought about talking about either the house itself, its door, its roof, something related to the house, or even a completely different subject. Let us consider the probability distribution of these alternatives (this notion of probability distribution is related, but is different from Steedman’s [2000] theme alternatives) as an event labeled as T for theme. We also consider the probability distribution for the alternatives for the rheme (labeled as R). There must be a variety of alternatives such as *large*, *wooden*, *flat*, *expensive*, and so on. It is important to realize that these probability distributions, T and R , are established before making the utterance in question.

How we can actually compute a probability distribution is a difficult question. Since some possibilities can be related to the context through inference, it naturally involves the kind of difficulty faced in many pragmatic studies. Next, there is a question such as whether the probability distribution under discussion should be understood only from the speaker’s point of view. In addition, the notion of joint entropy involves the connection between two events, which also requires analysis. For the present discussion, we assume that the probability distributions for the theme and the rheme are available and build arguments based on the assumption.

Now, we can think of the entropies for the theme and the rheme written as $H(T)$ and $H(R)$, respectively. The entire utterance has its own entropy, which is the joint entropy of the theme and the rheme, i.e., $H(T, R)$, independent of the theme-rheme ordering. In general, since there is some dependency between the theme and the rheme, the joint entropy is no greater than the sum of the entropies for the theme and rheme, i.e., $H(T, R) \leq H(T) + H(R)$. Since the rheme is pronounced after the theme, we consider the conditional entropy of the rheme after excluding the effect of the theme as $H(R|T)$. Then, $H(T, R) = H(T) + H(R|T)$. If the utterance is made in the rheme-theme order, we have $H(T, R) = H(R) + H(T|R)$, still with the same amount of the total (joint) information associated with the proposition.

Next, let us observe an example where the theme-rheme and the rheme-theme ordering differ with respect to the distribution of entropies. For simplicity, let us suppose that we have two possibilities for the theme and five possibilities for the rheme with certain dependency among them. In particular, we consider the probability distribution of the data (26) in Appendix A. For the both ordering, the relevant entropy measures are shown below.

(5) a.	Theme	Rheme	
	$H(T)$	$H(R T)$	Standard Deviation
	1.00	1.84	0.42
b.			
	Rheme	Theme	
	$H(R)$	$H(T R)$	Standard Deviation
	2.26	0.59	0.83

$H(T, R)$ is 2.84 for both cases. The above result shows that the theme-rheme order has more even distribution of entropies than the rheme-theme order. That is, it would be easier for the listener to process the information in the theme-rheme order. To measure the evenness of the entropies for the theme and the rheme, we compute the standard deviation of the two entropies as shown in (5). Note that we are not interested in whether there is any specific capacity of the listener or, if there is, what would be the value.

In order to apply the scheme in general, let us define the following term.

(6) (Definition) Information balance: The standard deviation of the entropies for the theme and the rheme for a particular ordering.

The main proposition of this paper is then described as follows:

(7) (Main Proposition) The information structure with a lower information balance is preferred.

Next, we consider the following theorem, which is proven in Appendix A.

(8) (Theorem) If the entropy of the theme is lower than that of the rheme, the theme-rheme ordering is never worse than the other ordering with respect to information balance.

I suggest that this is the source of theme-first tendency. The above theorem is interesting because of the following two points: (i) it predicts that the theme-rheme ordering is never worse than the other ordering and (ii) it can also specify under what condition there is no difference between the two ordering.

In certain cases, information balance can be the same for both the theme-rheme and the rheme-theme orderings. First, if the theme and the rheme are completely independent, the joint entropy is the sum of $H(T)$ and $H(R)$, i.e., $H(T, R) = H(T) + H(R|T) = H(T) + H(R)$. Thus, the ordering does not matter. But except for a special subcase discussed below, I suspect that this case rarely occur because an utterance is made to compose a meaningful proposition out of the theme and the rheme.

One important special case of the above is as follows. If the theme is completely predictable, i.e., $H(T) = 0$, the entire information solely depends on $H(R)$. The information balance is now between 0 and $H(R)$ regardless of the ordering. The situation corresponds to Lambrecht's [1994, p. 201] statement: if theme (his 'topic') is established, there is no necessity for it to appear sentence-initially. Although the symmetrical case where $H(R) = 0$ is theoretically possible, we do not consider it as we can assume that the rheme always have some information. One might wonder whether a redundant utterance (which is reported

to be common [Walker, 1992]) has no information at all, i.e., $H(T, R) = H(T) = H(R) = 0$. Although it is not crucial for the current discussion, it would be possible to maintain $H(R) > 0$. For example, a redundant utterance may ‘conversationally implicate’ some information in the sense of Grice [1975].

Another unlikely scenario is that the theme and the rheme are completely dependent. In such a case, the information balance would be again the same for the two ordering.

The consequences of the main proposition are as follows. Assuming that the theme has a lower entropy than the rheme, the theme-rheme ordering is never worse than the other with respect to information balance. Exceptions to theme-first principles occurs when the theme is completely predictable, i.e., $H(T) = 0$. We now turn to linguistic data to see whether this proposal is consistent with them.

3 Analysis of Rheme-First Cases

In this section, we examine various rheme-first cases. The first subsection deals with exceptions in English, which are not claimed to be rheme-first. The second subsection deals with examples in arguably systematically rheme-first languages.

3.1 Exceptions in English

Although we observed (1) as a rheme-first example, there are a few points we need to clarify. First, the verb *say* is an stage-level predicate (roughly, corresponding to a temporary state), not an individual-level predicate (roughly, corresponding to a permanent state), following the distinction of Carlson [1980]. Now, Kratzer [1995] argues that while stage-level predicates have an event argument, individual-level predicates do not. Then, the example (1) could be analyzed as follows:

- (9) *a.* Who said that (at that time)?
b. (At that time.)_{deleted-Theme} [**Peter**]_{Rheme} [said it]_{Theme}.

This type of event theme has been considered in the literature [Erteschik-Shir, 1998]. Then, the above may not be a good counterexample to theme-first principles as we might hypothesize a deleted theme. As in the above example, we accept that information-structure components, theme and rheme, be split into discontinuous sections. It is possible to analyze them, e.g., by adopting the idea of ‘structured meaning’ [Krifka, 1992].

To see whether a rheme-first example is possible in English, let us consider the following involving an individual-level predicate.

- (10) *a.* Who knows the secret?
b. [**Peter**]_{Rheme} [knows it]_{Theme}.

Since the individual-level verb *know* applies to a permanent state, no event argument is assumed. Thus, excluding the possibility of an extra event argument, it is still possible to have a rheme-first utterance in English. But as discussed in the previous section, this corresponds to the case where $H(T) = 0$. Thus, it is perfectly consistent with the present proposal.

Another point we need to discuss is the status of contrastive theme. Let us consider the following example from Jackendoff [1972, p. 261].

- (11) *Q:* Well, what about the **beans**? Who ate **them**?
A: [**Fred**]_{Rheme} [ate the **beans**]_{Theme}.

This is a stage-level example and the above-mentioned caveat applies. But we could create an individual-level counterpart, and thus, it does not really matter. Here, the word *beans* is stressed because of the potential contrast between the beans and, say, the potatoes. On the other hand, the predicate “ate the beans” is completely predictable at the time the response was uttered. Thus, the entropy of the predicate (theme) is 0.

In this respect, the possibilities to compute the entropy of a theme or a rheme is different from the instance of alternatives set as discussed in Steedman [2000]. The entropy of the theme does not necessarily depend on the contrastiveness of the theme. This type of rheme-theme ordering must be the result of the SVO syntax in English.

Lambrecht [1994, p. 202] argues that contrastive themes (his ‘topic’) must appear sentence-initially because they must announce a new topic or marking a topic shift. But as seen in (11), this statement is not correct. The present proposal differs from Lambrecht in that the predictability is computed independent of contrastiveness, and thus does not have the same problem.

To see the effect of word order difference, let us compare the following two short discourses.

- (12) *i.* Once upon a time, there were all kinds of vegetables in the field. There always were someone who ate some of them.
ii. Fred ate the beans.
iii. Fred was a monk who ...
- (13) *i.* Once upon a time, there were all kinds of vegetables in the field. There always were someone who ate some of them.
ii. The one who ate the beans was Fred.
iii. Fred was a monk who ...

It seems that (12*ii*) is no longer as good as (11A). (13*ii*) seems more natural.¹ Except for direct responses to a question, it would be generally difficult to fix the theme. This may be one of the reasons that we tend to observe mostly theme-first patterns in written discourse.

The present proposal predicts that it is preferable for an unpredictable theme to precede the rheme. But it is always possible that such preference be violated. As an example, consider the following abstract taken from a medical journal (sentences are numbered for reference purposes).

Title: ⁰Overuse Injuries in Children and Adolescents

¹The benefits of regular exercise are not limited to adults. ²Youth athletic programs provide opportunities to improve self-esteem, acquire leadership skills and self-discipline, and develop general fitness and motor skills. ³Peer socialization is another important, though sometimes overlooked, benefit. ⁴Participation, however, is not without injury risk. ⁵While acute trauma and rare catastrophic injuries draw much attention, overuse injuries are increasingly common.

⁶Diagnostic and treatment efforts should focus on how the injury developed and consider issues that are unique to growing athletes. ⁷An understanding of these concepts provides the basis for making specific injury-prevention recommendations.

The subject of Utterance 3, “peer socialization” might be inferrable from the context, but it seems more like a new concept. On the other hand, the predicate “another important, though sometimes overlooked, benefit” is more readily inferrable from “the benefits of regular exercise” in Utterance 1. Thus, it seems possible to analyze the utterance as in the rheme-theme order. However, this inferrable theme is by no means completely predictable from the context at the time of this utterance. As a result, this utterance does not conform to the hypothesis stated earlier. Now, we consider replacing Utterance 3 with the following: “Another important, though sometimes overlooked, benefit is peer socialization.” This seems more felicitous word order within this discourse.

To summarize, the counterexample to theme-first principles seem to fit the current proposal. More detailed examination of the applicability of the proposal remains as future work.

¹Discourse coherence of this type has been analyzed using Centering Theory Grosz et al. [1995]. It would be interesting to compare the present proposal with Centering Theory.

3.2 Rheme-First Languages

According to Lambrecht [1994, p. 200], the existence of verb-initial languages is a greater problem for theme-first principles. He acknowledges that it might be the case that every verb-initial language have topicalization. But such a construction is still a marked one compared to the basic ones. On the other hand, we need to be careful about identifying rheme-first patterns. First, depending on the way it is defined, typological classification of verb-initial language may simply mean that the pattern occurs more frequently than others. Second, being verb-initial does not automatically mean that the language is full of rheme-first patterns Payne [1995, p. 464]. The discussion below focuses on the data taken from Mithun [1995], which seems to represent the most prominently rheme-first case ('newsworthiness'-first, in her term).

We now examine Iroquoian data from Mithun [1995] (partially shown in Introduction). All the utterances shown below are taken from Tuscarora stories. The rheme indications are added for presentation purposes. The speaker first describes a long journey on the ice, discovery of land, and preparation for a sacrifice.

- (14) i. [ha? uhq?nq? ru?nq?qh]_{Rheme}, währáhrq?, ...
 the head man he said
 "the headman said, ..."
 ∴ (after the sacrifice is made)
- ii. q:waeh tihruyáhw?ah haent:kq: uhq?nq? ru?nq?qh?
 where he has learned from that head man
 "Where had he learned it, that headman?"
 ∴ (the speaker begins his recipe for cornbread)
- iii. Tyahraetšihq kq:θ [uhsaéharaeh]_{Rheme} ... wa?kkúhae?
 first customarily ash I went after
 "First, I usually would go after ashes."
 ∴ (after a kettle is prepared and is boiling)
- iv. U:nq kq:θ [yahwa?kkq?naé:ti?]_{Rheme} hä;thu ha?uhsaéharaeh.
 then customarily there I poured there the ash
 "Then I would pour the ashes in there."

We exclude the utterance (ii) from discussion because the information structure of a question is beyond the scope. First, (iii) and (iv) include an adverbial at the beginning of the utterance. Thus, it does not look strictly rheme-first in the sense of theme-first principles. On the other hand, the last constituent is a part of the theme in all utterances. Thus, the rheme-theme pattern is always present, and it is strikingly different from 'more' theme-first languages. We still want to show that the theme after the rheme is predictable. The constituents after the rheme are either a pronoun, a definite expression or a fairly light verb. We can say that they are highly predictable and their entropy are very low.

Let us examine additional utterances in Mithun. The following is an introductory sentence to begin a war story.

- (15) U:nqha? kyaent:kq: tikhà:wi? kyaent:kq: [kayq?ri:yus]_{Rheme}
 long ago this so it carries this they fight
 kyaent:kq: wahstqhá:ka:?, tinsnq? kuráhku:
 this Bostonians and British
 "One time long ago the Americans and the British were at war."

In this case, the sentence-initial constituents before “they fight” is actually a part of the theme and that seems to set the context. The English translation could be “One time long ago there was a war between the Americans and the British”.

In the following, a peddler has been driving a horse although the horse itself is not mentioned. Mithun argues that the newsworthiness of the verb.

- (16) *U:nq haésnq:* [θahra?nù:ri?]_{Rheme} ha?á:ha:θ.
 now then again he drove the horse

“Now then he drove his horse again.”

Again, the sentence-initial adverbial sets the event, which is a part of the theme.

Mithun does not discuss the context for the following, but says that the crucial point is “behind her”.

- (17) [*ae?taéhsnakw*]_{Rheme} wahra?ná?nihr.
 behind her he stood

“He stood behind her.”

Next, the main point is making fire.

- (18) [*Yú:naeks*]_{Rheme} uhá?nq?
 it burns in front

“A fire was burning before her.”

Mithun [1995, p. 391] cites the literature and observes that in spoken language, significant new ideas are introduced one at a time. In some examples above, we could even say that the story can continue by linking just the rhemes omitting the constituent after the rheme. Thus, these rheme-theme patterns too seem consistent with the present proposal.

Why there are (more or less) rheme-first languages and why there are so few are intriguing question. As a cognitive motivation for the rheme-first pattern, Downing [1995, p. 16] refers to ‘primacy effect’ [Gernsbacher and Hargreaves, 1992]. In addition, Mithun [1995, Sec. 4] adds that because of downstepping the sentence-initial position has an advantage of being more prominent (in absolute scale). However, since even Iroquoian allows sentence-initial adverbials as a part of the theme, neither of these proposals seem to apply in a strong form. Here is another question. In SOV languages, focus (of the rheme) most commonly appear on the immediately pre-verbal position [e.g., Kuno, 1978]. Use of such a position cannot be explained by primacy effect (first position) or recency effect (last position).

Next, the rheme-first languages discussed Mithun [1995] and others are highly agglutinating. As a result, the number of average constituents in an utterance seems smaller in such languages. Mithun [1995] explains the different degree of rheme-first tendency in the Siouan, Caddoan, and Iroquoian language in relation to the development of affixes. These factors may have some effects on the way the rheme is placed.

Additional relevant data can also be found in the following paper: Lambrecht [1987], Mithun [1987], Payne [1992], and Tomlin and Rhodes [1992]. These are left for future work.

4 On the Definition of Information Structure

So far we have been focusing on the discussion of theme-first tendency and exceptions assuming a general idea about information structure. In this section, we turn our attention to the definition of information structure itself. Note that this section provides an additional perspective on the assumption of the main point of this paper but must be considered separately from the main argument.

Although there are some general agreement about the notion of information structure, the precise definition is still a matter of controversy. This section adds yet another definition because it is rather different from the previous ones and might actually be combined with other definitions successfully.

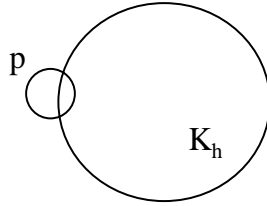


Figure 2: Vallduví's view of information update

4.1 Previous Definitions

Lambrecht [1994, p. 5] defines ‘information structure’ as follows:

- (19) That component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts.

This definition appears intuitive, but still does not nail down the concept in a precise manner. In particular, its reference to mental states seems to leave room for further specification.

As the basis for discussing information structure, Vallduví [1990, p. 15] refers to the notion of ‘information’ discussed by Dretske [1999]. A typical case of information update is represented as in Fig. 2. The figure suggests that a proposition (represented as p) has a component known to the hearer (inside the K_h circle) and another one that is new to the hearer (outside of it). Although the figure has some intuitive merits, its precise interpretation is not so straightforward. For example, where exactly the proposition p divided between inside and outside K_h ?

Now, let us consider the following example.

- (20) *Q*: Who did Felix praise?

A: [Felix praised]_{Theme} [himself]_{Rheme}.

The concept “Felix praised someone (or even nobody)” is in the shared knowledge. The referent Felix is also in the shared knowledge as pointed out by Reinhart [1982]. But the proposition (as a whole) is new. What makes a rheme as such is that the particular rheme is chosen in contrast to something else. Vallduví's [1990] schematic is not particularly helpful to analyze this point.

Although the referential status of the rheme can vary, there are certain restrictions on the referential status of the theme. Themes are in general ‘evoked’ or ‘inferable’ in the sense of Prince [1981]. However, it is extremely difficult to nail down to what extent we can actually infer a theme from the context. Any definition of information structure based on the referential status of the theme would face this problem.

4.2 Information-Theoretic Definition

One assumption we have been making is that the theme has lower entropy than the rheme. In this section, we attempt to define information structure based on this idea.

The definition we will consider here is as follows:

- (21) (Definition) The information structure of an utterance is a binary partition (composition) of the semantic representation of the utterance between theme and rheme such that the entropy of the rheme is greater than that of the theme.

Let us examine some of the prominent features of this definition. First, it assumes a binary partition (cf. the next section for the possibility of multiple partitions). I assume that partitions are those grammatically

feasible ones. For example, Steedman [2000] provides a basis for such partition based on the formalism of Combinatory Categorical Grammar.

The presence of binary partition requires that there are both theme and rheme, and is not compatible with the all-rheme pattern. This position is consistent with the argument of Erteschik-Shir [1998]. This is in contrast to Lambrecht [1994, p. 15] who acknowledges the existence of all-rheme utterances and argues that the topic of an all-rheme utterance is the speaker. This idea does not seem correct; one can say an potentially all-rheme utterance that is nothing to do with the speaker. Further discussion about the possibility of all-rheme utterances can be found in another paper of mine (in progress) “Focus Projection and Information Structure”.²

The definition (21) is only relative between theme and rheme, and does not directly refer to absolute properties of theme or rheme. As mentioned in Section 2, the computation of entropy would eventually depend on the analysis of inference. Thus, various problems of dealing with inference will not go away. However, there seems advantageous to abstract away from the difficulty with inference all in the computation of entropy.

Except for the binary partition requirement, the definition (21) does not refer to linguistic notion such as reference to a verb, argument/adjunct, and structure (cf. Sgall et al. [1986] and Lambrecht [1994, p. 16]). As a result, the definition can be applied robustly to any construction in any language. Since information structure is a complex phenomenon, there surely will be cases where fine tuning is required. But the existence of a base on which refinement can be made consistently must be a welcome result.

Since the definition (21) is based on entropy that evaluates to a numeric value, it can be compared with our own greyish judgment. In many cases, it appears difficult to analyze information structure, especially in a written text. A theory of information structure might actually need to fail gracefully in a difficult case. The present approach seems to allow such a possibility unlike previous definitions. Furthermore, the use of probability distribution would still allow us to assign small probabilities to unexpected outcomes. This can be adopted to account for unexpected options and indirect responses to a question.

Although the definition (21) is a relatively weak view of information structure, it is sufficient for the current purpose because everything we need in this paper can be derived from it.

Let us now turn to some potential problems and open questions in connection to the above definitions. First, Rochemont’s [1986, p. 52] defines two types of foci: contrastive and presentation (non-contrastive). One might wonder if a presentation focus can form a rheme that does not have an alternative, which might end up with a zero entropy. But this does not need to be the case; we can always consider “nothing” as an alternative to an object. Technically speaking, if we consider the power set of a single element, it would always include the empty set as an alternative [Büring, 1997, p. 40]. This is justifiable because we can respond to a *what*-question with *nothing* [Jackendoff, 1972, p. 246].

Another potential problem is how to compute entropy. Theoretically, it will remain a problem as already mentioned earlier. Practically, various approximation techniques may be applicable. For example, statistical language modeling as reviewed in Manning and Schütze [1999] is a popular area in Computational Linguistics.

Finally, we also leave the analysis of the information structure in a question as future work.

4.3 Communicative Dynamism

So far, we have been restricting ourselves to the binary-partition approach to information structure. But with the current approach, it could be extended to multiple partitions (if that makes sense). This would naturally connect to the idea of Communicative Dynamism of Firbas [1964].

²Available on-line at “<http://www.cis.upenn.edu/~komagata/papers.html>”.

I personally think that binary partition makes more sense as it can be observed in virtually all languages. Suppose that we consider languages like Iroquoian. The number of constituents is fewer than other languages. The word order within a word is rigidly fixed by morpho-syntax. Thus, there seems less use of CD in this type of languages. In general, it would be difficult to demonstrate multiple divisions universally.

5 Conclusion

This preliminary paper proposes a hypothesis that information structure is to even out the information load of the theme and the rheme (information balance). Assuming that the theme is the low-entropy component of an information structure, we show that placing the theme before the rheme is never worse than the other ordering in this respect. This is reflected by the theme-first tendency. Even though it is not obvious whether there is an absolute limit on the capacity of human communication channel, it might make sense to assume that we may use as little channel capacity as possible. I believe that this formulation provides a relatively precise starting point for discussion of information structure and word order.

One of a few exceptional cases is crucial for explaining rheme-first cases. That is, if the theme has a zero entropy, the theme-rheme ordering does not affect the information balance. I explore some examples mainly from arguably rheme-first languages. My analysis is that in these languages, the rheme-theme ordering is consistently made after the theme is well set up and thus the theme in such an utterance has very low entropy. The same idea applies to other exceptional cases, e.g., in English. As a consequence, the current proposal seems consistent with the general idea of theme-first principle and also with the apparent exceptions to the idea.

The paper also discusses the adequacy of the definition of information structure based on information theory. The assumption is that the low-entropy component of a binary partition of an utterance corresponds to the theme. In this connection, I note that a rheme is a requirement, and a theme (or a deleted theme) is also a requirement. This will guarantee that there always is a information-theoretic contrast between theme (including the deleted one) and the rheme.

The current proposal is to some extent consistent with many other proposals about the relation between word order and information structure. I do not think the prediction of the hypothesis radically deviate from the previous work. However, I believe that the proposal is novel in that it relates the phenomena directly with the notion of entropy, which is widely applied to various fields including linguistics. This approach also introduces a possibility of applying psycholinguistic/cognitive techniques to explore the idea. I think the current approach is the first to derive both theme-first tendency and seemingly exceptional cases from a single hypothesis. This is welcome as we can now view more diverse phenomena within a fewer principles.

Although the current proposal is entirely theoretic, there might be a way to apply the idea to Natural Language Processing. For example, text generation and readability analysis of certain languages may involve word order as an important property; some recent work includes Kruijff-Korbayová et al. [2000, Sec. 1.4.2] and Ratnaparkhi [2001]. Further, entropy is used to compute certain linguistic properties, albeit rather different ones from the topic of this paper, as reviewed in Manning and Schütze [1999, Sec. 2.2.7].

The most commonly used ways of analyzing information structure is to rely on the so-called ‘question test’. But the test cannot be used in indirect responses, neither in monologue texts. However, it seems perfectly reasonable to analyze the information structure in an indirect response. One approach would be to hypothesize that the indirect response is actually a direct response to a different question. But coming up with an appropriate hypothetical question for an utterance is basically the same problem as identifying the information structure of the indirect response. So, this approach simply sidesteps the problem. In this connection, analyzing information structure in a discourse in general is still a widely open problem. This is also related to the lack of precise definition of information structure. The proposed definition of information structure based on information theory may shed some light on this issue.

A Basic Information Theory

A.1 Introduction

This appendix provides a summary of basic information theory (A.1-A.3), an example of computing information balance (A.4) and the proof of the theorem (8) (A.5). The primary source of the discussion on information theory (A.1-A.3) is Reza [1994].³ A more compact summary of roughly the same coverage can be found in Manning and Schütze [1999, Ch. 2]. Furthermore, a more informal presentation is available in Dretske [1999, Ch. 1-2].

Information theory originates from the work of Nyquist and Hartley in the 1920s and was more widely introduced by Shannon in 1940s (references can be found in the cited work above). The main idea is to measure ‘information’ in terms of ‘entropy’, which is informally related to the notions such as: informativeness, randomness, uncertainty, unexpectedness, degree of surprise, disorder, and chaos. The notion of entropy has its root actually in physics (thermodynamics). Some direct applications of Shannon’s theory include communication theory and cryptography. In addition, the idea has also been applied to various fields including economics and linguistics.

A.2 Entropy: A Measure of Information

Let us suppose that there are n possible outcomes, x_1, \dots, x_n , that may occur with an equal probability (uniform distribution). The probability of any of these events is $1/n$. In this case, the ‘entropy’ (a measure of information) of this situation (probability distribution) is $\log_2 n$.

For example, for a choice between two distinct outcomes, the entropy is $\log_2 2 = 1$, where 1 bit is sufficient to distinguish the outcome.⁴ Similarly, for a choice between 8 distinct outcomes, $\log_2 8 = 3$, where 3 bits are sufficient. The more possibilities, the more information we have and the more storage we need to record the result. The range of entropy is between 0 (completely predictable) and ∞ (completely chaotic).

The use of logarithmic function is essential for us to be able to deal with information as an ‘additive’ property. It is also related to our psychological sensitivity that is generally considered logarithmic rather than linear (e.g., sound intensity vs. human perception). If we represent $\log_2 n$ in terms of probability, i.e., $p = \frac{1}{n}$, we can define the entropy function (for uniform distribution) $H_{uniform} : \mathbb{R} \rightarrow \mathbb{R}$ as a function of probability.

$$H_{uniform}(p) = \log_2 n = -\log_2 \left(\frac{1}{n} \right) = -\log_2 p \quad (101)$$

As a result, we generally use the form $-\log_2 p$ with a negative sign to represent the entropy involving a probability.

Let us generalize the definition of entropy so that we can measure the information of non-uniform probability distribution. Suppose that there are n possible outcomes, $[x_1, \dots, x_n]$, with the following probability distribution $[p_1, p_2, \dots, p_n]$, where p_i is the probability of x_i , i.e., the shorthand for $P(X = x_i)$ or $P(x_i)$. Naturally, all the probabilities must sum to the unity, i.e., $\sum_{i=1}^n p_i = 1$. The idea is to average the information for all the outcomes. For a particular outcome x_i , the (pointwise) entropy is $-\log_2 p_i$. We need to weigh

³The first three chapters (about 130 pages) of the book is a very good introduction including basic concepts in probability and some discrete mathematics. The book contains many concrete examples, discusses why the entropy formula must be the way it is, and is mathematically accurate. It is also inexpensive. One disadvantage is that it is rather old (originally published in 1961) and thus uses some old notation, e.g., $P\{x\}$ for probability.

⁴Although it is not essential for computing entropy, the base 2 is used as in computer science.

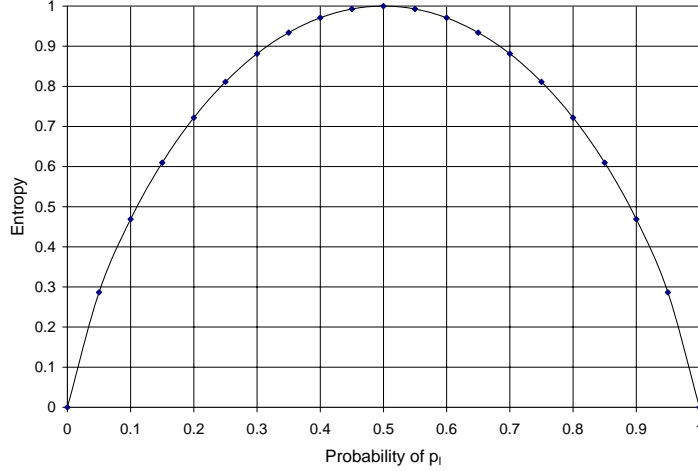


Figure 3: Entropy function for the binomial probability distribution

this value with its own probability, p_i . This leads to a term $p_i \log_2 p_i$, for x_i . We then add the weighted (pointwise) entropies for all the outcomes, resulting in the average of them (averaging makes sense due to the logarithmic conversion). Let us denote the probability distribution in question as \mathbf{p} (bold face to indicate that it is a vector, a ‘list’ of values). Then, the entropy $H : list(\mathbb{R}) \rightarrow \mathbb{R}$ can be represented as follows:

$$H(\mathbf{p}) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \cdots + p_n \log_2 p_n) = -\sum_{i=1}^n p_i \log_2 p_i \quad (102)$$

Although it is not obvious, Reza [1994, Secs 3-3 and 3-19] discusses that the entropy function is required to be in this particular form to satisfy the conditions: continuity, symmetry, and additivity.

It is straightforward to show that (101) is a special case of (102). If all $p_i = p$, the following holds, using $p = \frac{1}{n}$.⁵

$$H(\mathbf{p}) = -\sum_{i=1}^n p_i \log_2 p_i = -n \times p \log_2 p = -\log_2 p = H_{uniform}(p)$$

One of the most illustrative examples is a probability distribution with two distinct outcomes x_1 and x_2 . Naturally, $p_1 + p_2 = 1$. Thus, we have the following:

$$H(\mathbf{p}) = -\sum_{i=1}^2 p_i \log_2 p_i = -(p_1 \log_2 p_1 + p_2 \log_2 p_2) = -(p_1 \log_2 p_1 + (1 - p_1) \log_2 (1 - p_1))$$

This entropy function is shown in Fig. 3. The entropy is highest if the distribution is uniform, i.e., $p_1 = p_2 = 0.5$. This is because the likelihood of having one outcome is most uncertain. On the other extreme, if we know that x_1 (or x_2) always happens, the entropy is 0, i.e., there is no information.

⁵We often consider a random variable $X : \Omega \rightarrow \mathbb{R}$, where Ω is a sample space, and consider the entropy associated with it. In this case, we abuse the notation for the entropy function further and write $H(X)$. But the entropy is a function of a probability distribution and not of a random variable (function itself).

A.3 Information Measures for Two Events

Let us now consider two events X and Y . In the area of Communication Theory, X and Y are often viewed as the sender and the receiver (resp.) of signal. But later, we will interpret them as theme and rheme. Suppose that the event X has two possibilities x_1 and x_2 , and the event Y , two possibilities y_1 and y_2 . We now consider the probability ('joint probability') for each combination of x_i and y_j as follows:

(22)

	y_1	y_2
x_1	$p_{1,1}$	$p_{1,2}$
x_2	$p_{2,1}$	$p_{2,2}$

Naturally, the sum of all the probability must be exhaustive, i.e., $\sum_{j=1}^n \sum_{i=1}^m p_{i,j} = 1$.

At this point, we consider extending the definition of entropy (102) to this type of two-event situation. Instead of summing over a single event, we now sum over both of the events. For events X and Y with m and n possibilities, respectively, we have joint probability $p_{i,j}$ for x_i and y_j . Then, the entropy of the two events, 'joint entropy', is defined as follows:

$$H(X, Y) = - \sum_{j=1}^n \sum_{i=1}^m p_{i,j} \log_2 p_{i,j}$$

As an example, let us consider two events X and Y with the joint probability distribution as follows:

(23)

	y_1	y_2
x_1	0.1	0.2
x_2	0.3	0.4

Then, the joint entropy can be computed as follows:

$$H(X, Y) = - \sum_{j=1}^n \sum_{i=1}^m p_{i,j} \log_2 p_{i,j} = - (0.1 \log_2 0.1 + 0.2 \log_2 0.2 + 0.3 \log_2 0.3 + 0.4 \log_2 0.4) \simeq 1.84$$

Note that $H(X) \simeq 0.88$ and $H(Y) \simeq 0.97$. In this case, $H(X, Y) < H(X) + H(Y)$.⁶ Since the joint probability information already contains the complete information about the two events, knowing X and Y separately has some redundancy in terms of information. This situation is schematically shown as the diagram (a) in Fig. 4, where there is some overlap.

We now compare the above pattern with the two extreme cases. First, consider the following joint probability distribution:

(24)

	y_1	y_2
x_1	0.1	0.0
x_2	0.0	0.9

Then, the joint entropy is: $H(X, Y) = - (0.1 \log_2 0.1 + 0.9 \log_2 0.9) \simeq 0.47$. Note that $H(X) \simeq 0.47$ and $H(Y) \simeq 0.47$. In this case, $H(X, Y) = H(X) = H(Y)$. In fact, if non-diagonal entries (of a square matrix) are all 0, this equation always holds. If the information of X , Y , and the joint information of X and Y are all equal, we infer that they all have the same amount of information. That is, X and Y are completely

⁶More precisely, we have $H(X, Y) = 1.846439345$ and $H(X) + H(Y) = 1.852241494$. Thus, there is a small difference, 0.005802149.

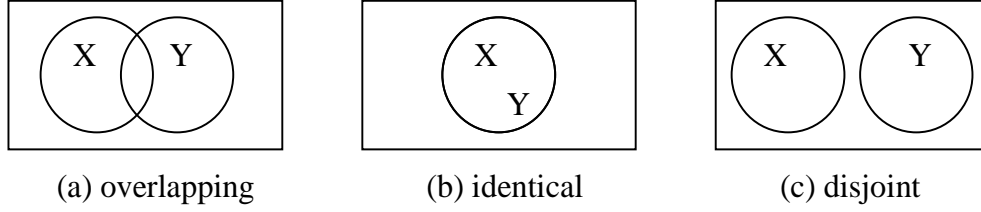


Figure 4: Relation between two events

dependent and knowing one of them is sufficient to know any of the other information measure. This situation is schematically shown as the diagram (b) in Fig. 4, where the two events completely overlap.

Next, consider the following example:

(25)

	y_1	y_2
x_1	0.1	0.1
x_2	0.4	0.4

Then, the joint entropy is: $H(X, Y) \simeq 1.72$. Note that $H(X) \simeq 0.72$ and $H(Y) \simeq 1.00$. In this case, $H(X, Y) = H(X) + H(Y)$. In fact, if the distributions are even across rows and columns, this equation always holds. In this case, the information about X and Y are completely *independent*. This situation is schematically shown as the diagram (c) in Fig. 4, where there is no overlap.

Except for the two extreme cases discussed above, it would be convenient to know where the difference between $H(X, Y)$ and $H(X) + H(Y)$ lies. The source of such difference is information dependency between X and Y as suggested in the example (23). At this point, let us consider the information measure that corresponds to $H(X, Y) - H(X)$. Schematically, this is shown as the shaded area in Fig. 1. Since this area Y is conditional to X , it is called ‘conditional entropy’, represented as $H(Y|X)$. Analogously, we can also consider $H(X|Y)$. Then, the following equations relate the information measures discussed so far.⁷

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Returning to the example (23), $H(X, Y) \simeq 1.84 = H(X) + H(Y|X) \simeq 0.88 + H(Y|X)$. Thus, we know

⁷Conditional entropy can be defined in terms of ‘conditional probability’, e.g., the probability of y_j when x_i is observed as shown below.

$$p_{j|i} = \frac{P(X = x_i \text{ and } Y = y_j)}{P(X = x_i)} = \frac{p_{i,j}}{P(x_i)}$$

Then, the conditional entropy of Y with the observation of x_i can be defined as the average over all the outcomes of Y as follows:

$$H(Y|x_i) = - \sum_{j=1}^n p_{j|i} \log_2 p_{j|i}$$

If we average over all the outcomes of X , the ‘conditional entropy’ of Y with X is defined as follows:

$$H(Y|X) = - \sum_{i=1}^m \sum_{j=1}^n p_{j|i} \log_2 p_{j|i}$$

that $H(Y|X)$ is 0.96, which is less than $H(Y) \simeq 1.97$. Since conditional information never increases the uncertainty, we have the following inequality.

$$H(X|Y) \leq H(X)$$

Another measure is used to indicate the degree of dependence between the two events. It is called ‘mutual information’ and characterized by the following equation:

$$I(X;Y) = H(X) + H(Y) - H(Y|X)$$

A.4 Information Balance

This subsection shows a computation of ‘information balance’ (6) introduced in Section 2. Let us consider the probability distributions for the theme and the rheme as the two events discussed in the previous subsection, denoted as T and R , respectively. We suppose that the theme has two possibilities t_1 and t_2 , and the rheme has five: r_1, \dots, r_5 , and the joint probability distribution is as follows.

(26)

	r_1	r_2	r_3	r_4	r_5	$\sum t_i$
t_1	0.25	0.125	0.075	0.025	0.025	0.5
t_2	0.025	0.025	0.075	0.125	0.25	0.5
$\sum r_i$	0.275	0.15	0.15	0.15	0.275	

If the word order is theme-rheme, we consider $H(T)$ and $H(R|T)$ because of the dependency observed in the given word order. The entropy of the entire utterance can be described as: $H(T, R) = H(T) + H(R|T)$. If the ordering is reversed, the entropy would be $H(T, R) = H(R) + H(T|R)$. Regardless of the ordering, the total information, i.e., the joint information, is identical as the same amount of information is eventually delivered.

The basic information measures are computed as follows:

$$\begin{aligned}
 H(T) &= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1.000 \\
 H(R) &= -(2 \times 0.275 \log_2 0.275 + 3 \times 0.15 \log_2 0.15) \simeq 2.256 \\
 H(T, R) &= -(2 \times 0.25 \log_2 0.25 + 2 \times 0.125 \log_2 0.125 + 2 \times 0.075 \log_2 0.075 + 4 \times 0.025 \log_2 0.025) \\
 &\simeq 2.843 \\
 H(R|T) &= H(T, R) - H(T) \simeq 1.843 \\
 H(T|R) &= H(T, R) - H(R) \simeq 0.587 \\
 I(T; R) &= H(T) + H(R) - H(T, R) \simeq 0.413
 \end{aligned}$$

Now, let us compute the average of the entropies for the theme and the rheme (same for the theme-rheme and the rheme-theme ordering).

$$E_{TR} = \frac{H(T) + H(R|T)}{2} = \frac{H(T) + H(T, R) - H(T)}{2} = \frac{H(T, R)}{2} = E_{RT} \simeq 1.421$$

Next, let us denote the information balance for the theme-rheme (rheme-theme) ordering as σ_{TR} (σ_{RT}). Then, the information balance for the two orderings can be computed as follows:

$$\begin{aligned}\sigma_{TR} &= \sqrt{\frac{|H(T) - E_{TR}|^2 + |H(R|T) - E_{TR}|^2}{2}} = \sqrt{\frac{|1.000 - 1.421|^2 + |1.843 - 1.421|^2}{2}} \simeq 0.421 \\ \sigma_{RT} &= \sqrt{\frac{|H(R) - E_{RT}|^2 + |H(T|R) - E_{RT}|^2}{2}} = \sqrt{\frac{|2.256 - 1.421|^2 + |0.587 - 1.421|^2}{2}} \simeq 0.835\end{aligned}$$

Thus, we have $\sigma_{TR} < \sigma_{RT}$.

A.5 Analysis of Information Balance

In this subsection, we prove the theorem (8) discussed in Section 2. As in the previous subsection, let us consider the entropies for T and R as $H(T)$ and $H(R)$, respectively. We also consider their joint entropy $H(T, R)$ and conditional entropies $H(R|T)$ and $H(T|R)$.

In general, the information balance for the two events X and Y in that ordering is computed as follows:⁸

$$\begin{aligned}\sigma_{XY} &= \sqrt{\frac{|H(X) - E_{XY}|^2 + |H(Y|X) - E_{XY}|^2}{2}} \\ &\quad \text{where } E_{XY} = \frac{H(X, Y)}{2} \text{ (as in the previous subsection)} \\ 2\sigma_{XY}^2 &= \left| H(X) - \frac{H(X, Y)}{2} \right|^2 + \left| H(X, Y) - H(X) - \frac{H(X, Y)}{2} \right|^2 \\ &= \left| H(X) - \frac{H(X, Y)}{2} \right|^2 + \left| -H(X) + \frac{H(X, Y)}{2} \right|^2 \\ \sigma_{XY}^2 &= \sqrt{\left| H(X) - \frac{H(X, Y)}{2} \right|^2} \\ \sigma_{XY} &= \left| H(X) - \frac{H(X, Y)}{2} \right|\end{aligned}$$

Now, we will prove the theorem, which is equivalent to the following:

(27) If $H(T) \leq H(R)$, $\sigma_{TR} \leq \sigma_{RT}$.

First, since $\frac{H(T, R)}{2} = E_{XY}$ and $H(T) < H(R)$,

$$\begin{aligned}\sigma_{TR} &= \left| H(T) - \frac{H(T, R)}{2} \right| = \frac{H(T, R)}{2} - H(T) \\ \sigma_{RT} &= \left| H(R) - \frac{H(T, R)}{2} \right| = H(R) - \frac{H(T, R)}{2}\end{aligned}$$

Then, applying $H(X, Y) = H(Y) + H(X|Y)$ and $H(X|Y) \leq H(X)$,

$$\begin{aligned}\sigma_{TR} - \sigma_{RT} &= \left[\frac{H(T, R)}{2} - H(T) \right] - \left[H(R) - \frac{H(T, R)}{2} \right] \\ &= H(T, R) - H(R) - H(T) \\ &= H(T|R) - H(T) \leq 0\end{aligned}$$

⁸The definition can be extended to multiple events.

Therefore, $\sigma_{TR} \leq \sigma_{RT}$.

As discussed in the paper, there are a few special cases. First, the completely independent case occurs when $I(X;Y) = 0$. Equivalently, $H(R|T) = H(R)$ and $H(X,Y) = H(X) + H(Y)$.

$$\sigma_{TR} = \left| H(T) - \frac{H(T,R)}{2} \right| = \frac{H(T,R)}{2}$$

$$\sigma_{RT} = \left| H(R) - \frac{H(T,R)}{2} \right| = \frac{H(T,R)}{2}$$

Therefore, there is no difference between the two ordering. As a special case, if $H(T) = 0$, we apply $H(T,R) = H(R)$ and obtain the following:

$$\sigma_{TR} = \frac{H(T,R)}{2}$$

$$\sigma_{RT} = H(R) - \frac{H(T,R)}{2} = \frac{H(T,R)}{2}$$

Again, there is no difference.

Next, let us consider the completely dependent case although this is not discussed in the paper (and unlikely in natural language as in the previous case). This case occurs when $H(X,Y) = H(X) = H(Y)$. Thus, there is no difference in this case either.

$$\sigma_{TR} = \left| H(T) - \frac{H(T,R)}{2} \right| = \frac{H(T,R)}{2}$$

$$\sigma_{RT} = \left| H(R) - \frac{H(T,R)}{2} \right| = \frac{H(T,R)}{2}$$

Bibliography

- Yehoshua Bar-Hillel. 1964. *Language and Information*. Addison-Wesley.
- Daniel Büring. 1997. The Great Scope Inversion Conspiracy. *Linguistics and Philosophy*, 20:175–194.
- Greg N. Carlson. 1980. *Reference to kinds in English* (originally a PhD thesis in 1977). Garland.
- Colin Cherry. 1978. *On human communication: a review, a survey, and a criticism*. MIT Press.
- Chet A. Creider and Jane T. Creider. 1983. Topic-Comment Relation in a Verb-Initial Language. *J. African Languages and Linguistics*, 5:1–15.
- Frederick J. Crosson and Kenneth M. Sayre, editors. 1967. *Philosophy and Cybernetics*. University of Notre Dame.
- Pamela Downing. 1995. Word order in discourse: By way of introduction. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*. John Benjamins.
- Fred I. Dretske. 1999. *Knowledge and the Flow of Information* (originally published in 1981 from the MIT Press). CSLI.
- Nomi Erteschik-Shir. 1998. The Syntax-Focus Structure Interface. In Peter W. Culicover and Louise McNally, editors, *Syntax and Semantics, Vol. 29: The limits of syntax*, pages 211–240. Academic Press.
- Jan Firbas. 1964. On Defining the Theme in Functional Sentence Analysis. *Travaux Linguistiques de Prague*, 1:267–280.

- Morton Ann Gernsbacher and David Hargreaves. 1992. The Privilege of Primacy: Experimental Data and Cognitive Explanations. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 83–116. John Benjamins.
- H. P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics, 3: Speech Acts*, pages 305–315. Academic Press.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–226.
- Michael A. K. Halliday. 1967. Notes on Transitivity and Theme in English (Part II). *Journal of Linguistics*, 3:199–244.
- Ray S. Jackendoff. 1972. *Semantic interpretation in Generative Grammar*. MIT Press.
- Otto Jespersen. 1924. *The Philosophy of Grammar*. London: Allen & Unwin.
- Angelica Kratzer. 1995. Stage-level and Individual-level Predicates. In Gregory N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 125–175. University of Chicago Press.
- Manfred Krifka. 1992. A Compositional Semantics for Multiple Focus Constructions. In Joachim Jacobs, editor, *Informationsstruktur und Grammatik (Linguistische Berichte, Sonderheft 4/1991-92)*, pages 17–53. Opladen: Westdeutscher Verlag.
- I. Kruijff-Korbayová, G.J.M. Kruijff, and John Bateman. 2000. Generation of Contextually Appropriate Word Order. In Kees van Deemter and Rodger Kibble, editors, *Information sharing*. CSLI.
- Susumu Kuno. 1978. *Danwa-no Bunpou (Discourse Grammar)*. Taishukan.
- Knud Lambrecht. 1987. On the status of SVO sentences in French discourse. In Russell S. Tomlin, editor, *Coherence and Grounding in Discourse*, pages 217–262. John Benjamins.
- Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, focus, and the mental representations of discourse referents*. Cambridge University Press.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundation of Statistical Natural Language Processing*. MIT Press.
- Vilém Mathesius. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis, edited by Josef Vachek*. The Hague: Mouton.
- Marianne Mithun. 1987. Is basic word order universal? In Russell S. Tomlin, editor, *Coherence and Grounding in Discourse*, pages 281–328. John Benjamins.
- Marianne Mithun. 1995. Morphological and prosodic forces shaping word order. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*. John Benjamins.
- Doris L. Payne. 1987. Information Structuring in Papago Narrative Discourse. *Language*, 63(4):783–804.
- Doris L. Payne. 1992. Nonidentifiable information and pragmatic order rules in ‘O’odham. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 137–166. John Benjamins.
- Doris L. Payne. 1995. Verb initial languages and information order. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*. John Benjamins.
- Ellen F. Prince. 1981. Toward a Taxonomy of Given-New Information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press.
- Adwait Ratnaparkhi. 2001. Modeling informational novelty in a conversational system with a hybrid statistical and grammar-based approach to natural language generation. In *NAACL Workshop on Adaptation*

- in Dialogue Systems, Pittsburgh, PA, June 2001.*
- Tanya Reinhart. 1982. Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, 27:53–94.
- Fazlollah M. Reza. 1994. *An Introduction to Information Theory (first published in 1961 by McGraw-Hill)*. Dover.
- Michael S. Rochemont. 1986. *Focus in generative grammar*. John Benjamins.
- Petr Sgall, Eva Hajičová, and Jarmila Panevova. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel.
- Mark Steedman. 2000. Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry*, 31(4):649–689.
- Russell S. Tomlin and Richard Rhodes. 1992. Information distribution in Ojibwa. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 117–136. John Benjamins.
- Enric Vallduví. 1990. *The informational component*. PhD thesis, University of Pennsylvania.
- Marilyn A. Walker. 1992. Redundancy in Collaborative Dialogue. In *COLING-92*, pages 345–351.