P. Blackburn et al. (Eds.): CONTEXT 2003, LNAI 2680, pp. 190-203, 2003.

Contextual Effects on Word Order: Information Structure and Information Theory

Nobo Komagata

Department of Computer Science, The College of New Jersey
PO Box 7718, Ewing, NJ 08628, USA
komagata@tcnj.edu

Abstract. To account for a type of contextual effect on word order, some researchers propose *theme-first* (old things first) principles. However, their universality has been questioned due to the existence of counterexamples and the possibility of arguably *rheme-first* (new things first) languages. Capturing the contextual effects on theme-rheme ordering (information structure) in terms of information theory, this paper argues that word order is affected by the distribution of informativeness, an idea also consistent with counterexamples and rheme-first languages.

1 Introduction

Various contextual effects on word order have been a topic of active research since at least the eighteenth century [1]. Many have noted that *old* information comes before *new* information [2–4]. The old and new components in an utterance are often called *theme* and *rheme*, respectively, and the theme-rheme organization is called *information structure*.[1] Accordingly, the idea of "old thing first" is also called the *theme-first principle*.

The theme-first principle seems to be able to account for certain word order phenomena, especially in *free-order* languages such as Czech [7]. Nevertheless, the proposal cannot be maintained in the stated form, because there are a number of counterexamples, such as the following. Note that bold face represents phonological prominence.

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

In the response in this example, the sentence-initial position is the rheme with new information corresponding to the *wh*-word in the question.

Furthermore, Lambrecht points out that a greater problem for the theme-first principle is the existence of arguably rheme-first languages [1]. For example, Mithun reports data from the Siouan, Caddoan, and Iroquoian languages and

[1] The contrast between theme and rheme is also referred to as the contrast between *topic* and *focus*, respectively. This paper uses the terms *theme* and *rheme*, focusing on the essence found in the contrast observed in many studies. Note that we assume that information structure is a binary partition at the utterance level [5, 6].

argues that these languages have a rheme-first tendency [8]. Similar data in other languages have also been reported [9–11]. Although it is not obvious that these are indeed rheme-first languages, the data still show a consistent pattern rather different from more theme-first languages.

Now that we cannot maintain the theme-first principle, at least as stated earlier, we must question whether something general can still be said about the contextual effects on word order in connection to information structure. 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. If different word order principles apply to different languages in an ad hoc way, it would pose a challenge to developing a universal account of language as a human cognitive process. Since information structure has been associated with word order in various forms, e.g., the Prague school [7] and strict theme-rheme ordering of Halliday [4], the above observation may undermine the role of information structure.

This paper develops an idea in Vallduví [12], who cites Dretske [13], regarding the notion of *information* (in terms of *entropy*) and analyzes word order from that point of view. In this connection, we also discuss the definition of information structure based on information theory.

The main hypothesis discussed here is that information structure is a means to even out the information load carried by the theme and the rheme of an utterance (referred to as *information balance*). Then, we can show that the ordering of a low-entropy theme followed by a high-entropy rheme is more desirable than the other ordering, which is considered the universal principle behind the theme-first tendency. However, if the theme is totally predictable (i.e., zero entropy), the ordering does not affect the information balance. This situation appears to correspond to apparent exceptions to the theme-first principle.

Word order is a complex phenomenon involving lexical, syntactic, and pragmatic constraints [14]. This paper inevitably leaves out certain important aspects, such as word order within a phrase, where morpho-syntax tends to fix word order quite rigidly.

The rest of this paper is organized as follows. Section 2 introduces an analysis of the theme-first principle based on information theory. Section 3 discusses various rheme-first cases and analyzes whether they are accountable within the current approach. Section 4 presents an information-theoretic definition of information structure.

2 Information-Theoretic Analysis of Word Order

In this section, we discuss the idea of applying information theory to the analysis of the theme-first principle using the following short discourse, where the second utterance is partitioned into a theme and a rheme.

(2) *i.* John has a house.
ii. [The **door**]$_{Theme}$ [is **purple**]$_{Rheme}$.

Compared to the above example, the following alternative appears less natural.

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

ii. [**Purple**]$_{Rheme}$, [the **door** is]$_{Theme}$.

The difference will be analyzed later in this section.

2.1 Basic Entropy Computation

This subsection discusses a way to compute the entropies of a theme and a rheme as independent events, using example (2). Immediately after the first utterance in the example, the speaker might want to talk about either the roof or the door, something related to the house, or even a completely different subject. For each of these subjects, there may be a variety of possible predicates, e.g., large, wooden, flat, expensive, purple, and so on. Although it is possible to demonstrate the computation of entropy for an arbitrarily complicated case, we use the following simplified scenario for presentation purposes: two choices for the theme between the door and the roof, and five choices for the rheme among yellow, red, orange, pink, and purple.

Roughly speaking, with more choices, the likelihood of choosing a particular option is smaller. In other words, the informativeness of a single choice among many would be higher than the one from fewer choices. This idea can be formally represented using the notion of *entropy* (good introductions include [15, 16]). Informally, high entropy is associated with high informativeness, low predictability, high uncertainty, more surprise, etc. The use of entropy has been discussed even in linguistics and philosophy [17–19]. For example, while Cherry suggests usefulness [19], Bar-Hillel is more cautious, saying that *information* is different from *meaning* [17]. Naturally, the focus of this paper is not on meaning, but on word order.

Under a very special case where all the events are equally likely (uniform distribution), the entropy of an 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, but it is also adjusted logarithmically so that it is *additive*, in accordance to human sense. For n equally-likely outcomes, $x_1, ..., x_n$, the entropy is defined as a function $H_{uniform}$ on real numbers:[2]

$$H_{uniform}\left(p\right) = \log_2 n = -\log_2\left(1/n\right) = -\log_2 p \ .$$

For example, under the current scenario for example (2), the entropy of the theme with two choices is $\log_2 2 = 1.0$, and the entropy of the rheme with five choices is $\log_2 5 \simeq 2.322$.

Entropy is a general function that can also be applied to an event X with n outcomes $[x_1, ..., x_n]$ and the corresponding probability distribution $[p_1, p_2, ..., p_n]$. Here, p_i is the probability of x_i, i.e., the shorthand for $P\left(X = x_i\right)$ or $P\left(x_i\right)$. Naturally, we must have $\sum_{i=1}^{n} p_i = 1$. For a particular outcome x_i, the (pointwise)

[2] The use of base 2 is convenient as it enables us to measure entropy in terms of *bit*.

entropy is $-\log_2 p_i$. We now compute the weighted average of the information for all the outcomes. That is, we multiply the ith entropy with its own probability, p_i, and then add them all (averaging makes sense due to the logarithmic conversion). Let us denote the probability distribution in question as $\mathbf{p}$ (bold face represents a vector, a *list* of values). Then, the entropy function H is defined as follows:

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

For example, if the five choices of the rheme in example (2) have a probability distribution $\mathbf{r} = [0.275, 0.15, 0.15, 0.15, 0.275]$, the entropy $H(\mathbf{r})$ is $-(2 \times 0.275 \log_2 0.275 + 3 \times 0.15 \log_2 0.15) \simeq 2.256$.

2.2 Dependency between Two Events

Although the entropies for a theme and a rheme were assumed independent in the previous subsection, the choice of the latter component would naturally depend on the choice of the former. For instance, in example (2), the predicates for the roof and those for the door are likely to have different probability distributions. In order to analyze the dependency between theme and rheme, this subsection introduces some basic ideas about entropies of two events.

We now consider two events X and Y. Suppose that event X has two possibilities, x_1 and x_2, and event Y, two possibilities y_1 and y_2. Then, the *joint probability* for each combination of x_i and y_i can be summarized as follows:

(5)

	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 probabilities must satisfy: $\sum_{j=1}^{n} \sum_{i=1}^{m} p_{i,j} = 1$.

At this point, we consider extending the definition of entropy (4) to a two-event situation, summing 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 *joint entropy* of the two events 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 the following joint probability distribution for X and Y:

(6)

	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:

$$H(X, Y) = -(0.1 \log_2 0.1 + 0.2 \log_2 0.2 + 0.3 \log_2 0.3 + 0.4 \log_2 0.4) \simeq 1.846\ .$$

Since the joint probability already contains the complete information about the two events, knowing X and Y separately would generally lead to some redundancy. For example, since $H(X) + H(Y) \simeq 0.881 + 0.971 = 1.852$, we see that $H(X,Y) < H(X) + H(Y)$.

We now consider the information measure that corresponds to $H(X,Y) - H(X)$. Since Y is conditional to X, it is called the *conditional entropy*, represented as $H(Y|X)$. Analogously, we can also consider $H(X|Y)$. Then, the following equation relates the information measures discussed so far.

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

Returning to example (6), we have $H(X,Y) = H(X) + H(Y|X) \simeq 0.881 + H(Y|X)$. Thus, we know that $H(Y|X)$ is 0.965, which is less than $H(Y) \simeq 0.971$. 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 two events, called *mutual information*, which is defined by the following equation: $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

2.3 Information Balance

We now apply the ideas introduced in the previous subsections to our analysis of word order. We use example (2) with the following probability distribution for the theme and the rheme (t_1 and t_2 refer to the two theme choices and r_i refers to one of the five rheme choices).

(7)

	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	

How we can actually come up with such a probability distribution is a difficult question. Since some possibilities can be related to the context through inference (linguistic and extra-linguistic), it naturally involves the kind of difficulty faced in many pragmatic studies. Next, there is a question of 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 we will build arguments based on this assumption.

The entropies for the theme, the rheme, and the entire utterance (independently) are $H(T)$, $H(R)$, and $H(T,R)$, respectively. If the rheme is delivered after the theme, we consider the conditional entropy of the rheme after excluding the effect of the theme, i.e., $H(R|T)$. Then, $H(T,R) = H(T) + H(R|T)$. On the other hand, if the utterance is made in the rheme-theme order, we have $H(T,R) = H(R) + H(T|R)$. In the following, as the entropy of the latter component, be it the rheme or the theme, we always use the conditional entropy. The basic information measures for example (7) are computed as follows:

$$
\begin{aligned}
H(T) &= -\left(0.5\log_2 0.5 + 0.5\log_2 0.5\right) = 1.000 \\
H(R) &= -\left(2\times 0.275\log_2 0.275 + 3\times 0.15\log_2 0.15\right) \simeq 2.256 \\
H(T,R) &= -(\; 2\times 0.25\log_2 0.25 + 2\times 0.125\log_2 0.125 \\
&\qquad +2\times 0.75\log_2 0.75 + 4\times 0.025\log_2 0.025) \simeq 2.843 \\
H(R|T) &= H(T,R) - H(T) \simeq 1.843 \\
H(T|R) &= H(T,R) - H(R) \simeq 0.587 \\
I(T;R) &= H(T) + H(R) - H(T,R) \simeq 0.413\ .
\end{aligned}
$$

In order to compare the evenness of the information distribution between theme and rheme, we introduce a measure, *information balance*, defined as follows:

Definition 1. *Information balance: The standard deviation of the entropies of the theme and the rheme (of an utterance) for a particular ordering.*

Note again that the entropy of the latter component is a conditional entropy. With this definition, the main proposition of this paper can be described as follows:

Proposition 1. *The information structure with a lower information balance is preferred.*

Next, let us compute the information balance of the theme-rheme (rheme-theme) ordering, denoted as σ_{TR} (σ_{RT}). To do so, we first compute the average of the entropies for the theme and the rheme (identical for both orders): $E_{TR} = E_{RT} = H(T,R)/2 \simeq 1.421$.

$$
\begin{aligned}
\sigma_{TR} &= \sqrt{\left(|H(T) - E_{TR}|^2 + |H(R|T) - E_{TR}|^2\right)/2} \simeq 0.421 \\
\sigma_{RT} &= \sqrt{\left(|H(R) - E_{RT}|^2 + |H(T|R) - E_{RT}|^2\right)/2} \simeq 0.835
\end{aligned}
$$

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

For both of the word orders, the relevant entropy measures and information balances are summarized below.

(8)	*a.*	Theme	Rheme	Information Balance
		$H(T)$	$H(R|T)$	σ_{TR}
		1.000	1.843	0.421
	b.	Rheme	Theme	Information Balance
		$H(R)$	$H(T|R)$	σ_{RT}
		2.256	0.587	0.835

This shows that the theme-rheme order has a 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.

Now, we can formulate the principle underlying the theme-first tendency as follows:

Theorem 1. *(Informally) 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. (Formally) If* $H(T) \leq H(R)$, $\sigma_{TR} \leq \sigma_{RT}$.

The above theorem is interesting on the following two points: (i) it predicts that the theme-rheme ordering is preferred, and (ii) it can also specify under what condition there is no difference between the two orders. Here is a proof.

Proof. First, the information balance for the two events X and Y in that ordering is computed as follows:

$$\begin{aligned}\sigma_{XY} &= \sqrt{\left(|H(X) - E_{TR}|^2 + |H(X,Y) - H(X) - E_{TR}|^2\right)/2} \\ &= \sqrt{\left(|H(X) - E_{TR}|^2 + |-H(X) + E_{TR}|^2\right)/2} \\ &= |H(X) - E_{TR}| \ .\end{aligned}$$

Let us consider the (independent) entropies for T and R as $H(T)$ and $H(R)$, respectively. Since $H(T) \leq H(R)$, we have $\sigma_{TR} = |H(T) - E_{TR}|$ and $\sigma_{RT} = |H(R) - E_{TR}|$. Then, applying $H(X,Y) = H(Y) + H(X|Y)$ and $H(X|Y) \leq H(X)$, we have the following.

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

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

2.4 Special Cases

As suggested in the previous subsection, information balance can be the same for both the theme-rheme and rheme-theme orders in certain cases.

First, the theme and the rheme could have exactly the same information (or are completely dependent), i.e., $H(T,R) = H(T) = H(R)$. However, this case is unlikely in reality.

Second, if the theme and the rheme are completely independent, i.e., $I(X;Y) = 0$, 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 information balance would not depend on the theme-rheme ordering. As we noted earlier, it is more likely that the theme and the rheme have some informational dependency, and thus this case would be atypical. However, there is an important special subcase. If the theme is completely predictable, i.e., $H(T) = 0$, the entire information solely depends on $H(R) = H(T,R)$, i.e., $\sigma_{TR} = \sigma_{RT}$. The information balance is now between zero and $H(R)$ regardless of the word order. The situation corresponds to Lambrecht's statement: if theme (his topic) is established, there is no need for it to appear sentence-initially [1]. The symmetrical case where $H(R) = 0$ is unlikely because we can assume that the rheme always has some information.

In summary, 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 the theme-first principle occur when the theme is completely predictable, i.e., $H(T) = 0$.

3 Analysis of Rheme-First Cases

In this section, we examine various rheme-first cases. The first subsection deals with exceptions in English, a language that is not considered rheme-first. The second subsection deals with examples in an arguably rheme-first language.

3.1 Exceptions in English

In example (1), the theme is completely predicable. Thus, its entropy is zero. As a result, it falls into the special case discussed in the previous section, where the position of the theme does not affect the information balance. Exceptions to strict theme-principles like this are still consistent with the present hypothesis.

There is another point regarding the status of contrastive theme, as in the following example.[3]

(9) *Q*: Well, what about the **beans**? Who ate **them**?

A: [**Fred**]$_{Rheme}$ [ate the **beans**]$_{Theme}$.

Here, the word "beans" is stressed because of the potential contrast between beans and, say, potatoes. One might question whether the entropy of such a theme is zero. But as long as the theme is completely predictable as in the above example, its entropy is still zero. Thus, the above example is consistent with our analysis. The existence of contrastive elements does not necessarily increase the entropy. In this respect, entropy computation is different from analyzing the set of alternatives as discussed in Steedman [22].

Lambrecht argues that contrastive themes (his topic) must appear sentence-initially because they must announce a new topic or mark a topic shift [1]. But example (9) is a counterexample to this analysis. Unlike Lambrecht, the present hypothesis predicts and accepts the existence of a contrastive theme after the rheme as long as it has zero entropy.

In written texts in English, it is generally more difficult to find a rheme-first pattern. Here is an attempt to create a text comparable to example (9).

(10) *i.* Once upon a time, the villagers planted beans and potatoes. One day, they noticed that someone ate the beans. Someone must have ate them.

ii. Fred ate the beans.

iii. Fred was a monk who ...

Although utterance (10*i*) provides basically the same information as question (9*Q*), utterance (10*ii*), which is the same as (9*A*), sounds less natural in this

[3] Predicates like "eat" imply the existence of a (possibly deleted) event argument [20], which may affect the information-theoretic analysis [21]. This situation can be avoided by using another type of verb, such as "know."

text. An alternative, "the one who ate beans was Fred," sounds more natural. This suggests that the entropy of "ate the beans" is not zero. Unlike the context generated by a question, utterances in a written text tend to leave a variety of options after them. This seems to explain why the theme-first tendency is observed more commonly in the written form of English.

The present hypothesis predicts the following: it is preferable for an unpredictable theme to precede a rheme. However, it is always possible to violate such a preference. As an example, consider the following abstract taken from a medical journal (utterances are numbered for reference purposes).

> Title: [0]Overuse Injuries in Children and Adolescents
>
> [1]The benefits of regular exercise are not limited to adults. [2]Youth athletic programs provide opportunities to improve self-esteem, acquire leadership skills and self-discipline, and develop general fitness and motor skills. [3]Peer socialization is another important, though sometimes overlooked, benefit. [4]Participation, however, is not without injury risk. [5]While acute trauma and rare catastrophic injuries draw much attention, overuse injuries are increasingly common.

In utterance 3, between the phrases (A) "peer socialization" and (B) "another important, though sometimes overlooked, benefit," phrase (B) seems to connect to the context more strongly due to the word "benefit," which already appeared in utterance 1. While the choice of "benefit" is among other contextually linked alternatives, the choice of "peer socialization" is among more diverse possibilities. Then, the entropy of phrase (B) must be lower than that of phrase (A). If the phrases are reversed as in "Another important, though sometimes overlooked, benefit is peer socialization," the information balance of this utterance would be lower and more appropriate than the original utterance 3 in this context.

3.2 Rheme-First Languages

Although some have claimed certain languages to be rheme-first, 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 [23].

The discussion below focuses on Iroquoian data taken from Mithun [8], which seems to represent the most prominently rheme-first case (*newsworthiness*-first, to use her term). The utterances are taken from Tuscarora stories. The background is as follows: the speaker first describes a long journey on the ice, discovery of land, and preparation for a sacrifice (some phonetic symbols have been replaced for font availability reasons: "*ą*" for right-hooked schwa and "*ʔ̱*" for glottal stop).

(11) *i.* [*haʔ̱ uhą́ʔ̱nąʔ̱ ruʔ̱ną́ʔ̱ąh*]$_{Rheme}$, *wahrą́hrąʔ̱, ...*
the head man he said
"the headman said, ..."

⋮ (after the sacrifice is made)

ii. *ą̀:waeh tihruyą́hwʔ̱ąh haení:ką: uhą́ʔ̱nąʔ̱ ruʔ̱ną́ʔ̱ąh?*
where he has learned from that head man
"Where had he learned it, that headman?"

⋮ (the speaker begins his recipe for cornbread)

iii. *Tyahraetšíhą ką:θ* [*uhsaéharœh*]$_{Rheme}$ ... *waʔ̱kkúhaeʔ̱.*
first customarily ash I went after
"First, I usually would go after ashes."

⋮ (after a kettle is prepared and is boiling)

iv. *U:ną ką:θ* [*yahwaʔ̱kkąʔ̱naé:tiʔ̱*]$_{Rheme}$ *hä¡thu haʔ̱uhsaéharaeh.*
then customarily there I poured there the ash
"Then I would pour the ashes in there."

We exclude utterance (*ii*) from discussion because analysis of the information structure of a question is beyond the scope of this paper. First, (*iii*) and (*iv*) include an adverbial at the beginning of the utterance. Thus, they do not have rheme-first patterns in a strict sense. On the other hand, the last constituent is a part of the theme in each utterance. Thus, we see some type of rheme-theme pattern consistently, which is strikingly different from *more* theme-first languages. The constituents after the rhemes are either a pronoun, a definite expression, or a fairly light verb. That is, these constituents are highly predictable and their entropies are very low, if not zero.

Let us examine other utterances from the same story. The following is an introductory sentence to begin a war story.

(12) *U:nąhaʔ̱ kyaení:ką: tikahà:wiʔ̱ kyaení:ką:* [*kayąʔ̱rì:yus*
long ago this so it carries this they fight
kyaení:ką: wahstąhá:ka:ʔ̱, tisnąʔ̱ kuráhku:]$_{Rheme}$.
this Bostonians and British
"One time long ago the Americans and the British were at war."

This is in fact a theme-rheme pattern. The theme is a typical element used to begin a story. The verb-subject order within the rheme is beyond the scope of the present analysis.

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

(13) *U:ną haésną:* [*θahraʔ̱nù:riʔ̱*]$_{Rheme}$ *haʔ̱á:ha:θ.*
now then again he drove the horse
"Now then he drove his horse again."

Again, this is not strictly rheme-first, and the constituent "the horse" is predictable from the context.

Mithun does not discuss the context for the following, but says that the focal point is "behind her."

(14) [*aeʔ̱taéhsnakw*]$_{Rheme}$ *wahraʔ̱náʔ̱nihr.*
behind her he stood
"He stood behind her."

The information “he stood” must be predictable. In the following, although “in front” is probably not completely predictable, it seems to have a low entropy readily inferrable from “behind.”

(15) [*Yú:ʔ̱naeks*]$_{Rheme}$ *uhą́ʔ̱nąʔ̱.*
it burns in front
“A fire was burning before her.”

Mithun cites the literature and observes that in spoken language, significant new ideas are introduced *one at a time* [8]. For the above example, we could even say that the story can continue by linking the rhemes (and the themes preceding the rhemes), but omitting the constituents after the rhemes. Thus, in these rheme-theme patterns, we can still see zero-entropy themes after the rhemes. This observation is consistent with the present hypothesis.

Why there are (more or less) rheme-first languages and why there are also so few are intriguing questions. As a cognitive motivation for the rheme-first pattern, Downing refers to *primacy effect* [24]. In addition, Mithun adds that the sentence-initial position has an advantage of being more prominent prosodically because of *downstepping* (gradually decreasing pitch) [8]. However, since even Iroquoian allows sentence-initial adverbials as a part of the theme, neither of these proposals seems convincing. Finally, Mithun points out that the arguably rheme-first languages are highly agglutinating with a small number of constituents in each utterance and that the development of affixes may have affected the different degree of rheme-first tendency in the Siouan, Caddoan, and Iroquoian languages [8]. Additional relevant data can also be found in the literature [25–28], which are left for future work.

4 On the Definition of Information Structure

In this section, we turn our attention to the definition of information structure. Although researchers have 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 could provide a precise foundation for its predecessors.

4.1 Previous Definitions

The most common way of analyzing information structure is to use a *question test*, as already seen in example (1). We could even define information structure based on a question test. However, such a definition cannot be applied to analyze information structure in texts. Another popular definition by Halliday [4] is problematic, because it is limited to the theme-rheme order.

Lambrecht provides a more general definition as shown below [1].

> *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 it 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.

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 *inferrable* in the sense of Prince [29]. However, it is extremely difficult to pinpoint 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 of our assumptions is that the theme has lower entropy than the rheme. In this section, we attempt to define information structure based on this idea. Here is our definition:[4]

Definition 2. *The information structure of an utterance is the linguistic realization of 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. We also assume that partitions are those grammatically feasible ones. For example, such a partition can be represented using Combinatory Categorial Grammar as discussed in Steedman [22].

Definition 2 refers to the entropies of the theme and the rheme only relatively and does not directly refer to absolute properties of the theme or the 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, it seems advantageous to abstract away from the difficulty with inference, as we can leave it all in the computation of entropy.

Except for the binary partition requirement, Definition 2 does not refer to linguistic notions such as reference to a verb and argument-adjunct distinction (cf. [7, 1]). As a result, the definition can be applied robustly to any construction in any language.

Since Definition 2 is based on entropy that evaluates to a numeric value, it can be compared with our own occasionally grayish judgment about information structure. In many cases, it is difficult to analyze information structure, especially in a written text. A theory of information structure may actually need to fail gracefully if the situation is not clear-cut. Unlike previous definitions, the present approach accepts such a possibility. 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.

[4] This definition is not compatible with recursive analyses of information structure including [4]. More details on this point are available in [6].

5 Conclusion

This 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, in this respect, never worse than the other order. A natural consequence is the theme-first tendency. One interpretation is that information structure is a way to minimize the required channel capacity.

The rheme-first examples are analyzed as involving zero-entropy themes. Since the information balance is not affected by the position of such themes, these examples are still consistent with our proposal. The paper also discusses a new definition of information structure as informational contrast between theme and rheme, which can serve as the basis for the entire discussion of this paper.

The current proposal is to some extent consistent with many other proposals about the relation between word order and information structure. However, the proposal is novel in that it relates certain word-order 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 for further evaluation. The proposal is arguably the first to derive both theme-first tendency and seemingly exceptional cases from a single hypothesis. This is desirable as we can now view more diverse phenomena with fewer principles.

References

1. Knud Lambrecht. *Information Structure and Sentence Form: Topic, focus, and the mental representations of discourse referents.* Cambridge University Press, 1994.
2. Vilém Mathesius. *A Functional Analysis of Present Day English on a General Linguistic Basis, edited by Josef Vachek.* The Hague: Mouton, 1975.
3. Jan Firbas. On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, 1:267–280, 1964.
4. M. A. K. Halliday. *An Introduction to Functional Grammar.* London: Edward Arnold, 1985.
5. Nobo N. Komagata. *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese.* PhD thesis, University of Pennsylvania, 1999.
6. Nobo Komagata. Information structure in subordinate and subordinate-like clauses in special issue on discourse and information structure (to appear). *Journal of Logic, Language and Information*, 12(3), 2003.
7. Petr Sgall, Eva Hajičová, and Jarmila Panevova. *The meaning of the sentence in its semantic and pragmatic aspects.* D. Reidel, 1986.
8. Marianne Mithun. Morphological and prosodic forces shaping word order. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse.* John Benjamins, 1995.
9. Doris L. Payne. Information structuring in papago narrative discourse. *Language*, 63(4):783–804, 1987.

10. Chet A. Creider and Jane T. Creider. Topic-comment relation in a verb-initial language. *J. African Languages and Linguistics*, 5:1–15, 1983.
11. Elena M. Leman. *Cheyenne Major Constituent Order.* Dallas, TX: Summer Institute of Linguistics, 1999.
12. Enric Vallduví. *The informational component.* PhD thesis, University of Pennsylvania, 1990.
13. Fred I. Dretske. *Knowledge and the Flow of Information (originally published in 1981 from the MIT Press).* CSLI, 1999.
14. I. Kruijff-Korbayová, G.J.M. Kruijff, and John Bateman. Generation of contextually appropriate word order. In Kees van Deemter and Rodger Kibble, editors, *Information sharing.* CSLI, 2000.
15. Fazlollah M. Reza. *An Introduction to Information Theory (first published in 1961 by McGraw-Hill).* Dover, 1994.
16. Christopher D. Manning and Hinrich Schütze. *Foundation of Statistical Natural Language Processing.* MIT Press, 1999.
17. Yehoshua Bar-Hillel. *Language and Information.* Addison-Wesley, 1964.
18. Frederick J. Crosson and Kenneth M. Sayre, editors. *Philosophy and Cybernetics.* University of Notre Dame, 1967.
19. Colin Cherry. *On human communication: a review, a survey, and a criticism.* MIT Press, 1978.
20. Angelica Kratzer. 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, 1995.
21. Nomi Erteschik-Shir. 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, 1998.
22. Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689, 2000.
23. Doris L. Payne. Verb initial languages and information order. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse.* John Benjamins, 1995.
24. Pamela Downing. Word order in discourse: By way of introduction. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse.* John Benjamins, 1995.
25. Knud Lambrecht. On the status of SVO sentences in French discourse. In Russell S. Tomlin, editor, *Coherence and Grounding in Discourse*, pages 217–262. John Benjamins, 1987.
26. Marianne Mithun. Is basic word order universal? In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 15–62. John Benjamins, 1992.
27. Doris L. Payne. Nonidentifiable information and pragmatic order rules in 'o'odham. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 137–166. John Benjamins, 1992.
28. Russell S. Tomlin and Richard Rhodes. Information distribution in Ojibwa. In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 117–136. John Benjamins, 1992.
29. Ellen F. Prince. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, 1981.