

Chapter 2

Information Structure: The State of the Art and Open Questions

In this chapter, we review existing theories of information structure and computational approaches to identifying information structure. We first point out that some existing definitions of information structure fail to explicate the properties of its components and the relation between the components. The next point is that most theoretical proposals about information structure are indifferent to the Identification Problem and lack the explicitness required for formalization and implementation. Finally, we observe that existing computational approaches do not yet provide a solution to the Identification Problem due to their limited coverage, lack of evaluation, and missing connection to theories.

To clarify our goal, we begin this chapter with a discussion of the Identification Problem for information structure. After presenting an informal view of information structure, we move to the review of theoretical and computational proposals in that order.

2.1 The Identification Problem

In the Introduction, we noted that the Identification Problem for information structure is necessary for applications such as machine translation, speech generation, and computer-assisted writing. This section explores this problem more in detail and identifies the associated subgoals.

The **Identification Problem** takes the following form. Given a text such as the one shown

below, the information structure consisting of two components, say, ‘theme’ and ‘rheme’, for each utterance except for the title must be identified (the text is taken from our experiment data, which will be discussed in Chapter 7).

(11) Title: Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

Osteoporosis has been defined as “a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk.” Although anyone can develop osteoporosis, postmenopausal women and young females with menstrual irregularities are most commonly affected. An estimated 20% of women more than 50 years old have osteoporosis. Although most studies have focused on women of this age-group, osteoporosis is potentially more deleterious in younger women because they haven’t yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives.

Now, suppose that a hypothetical procedure identifies the information structures as follows:

(12) Title: Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

[Osteoporosis]_{Theme} [has been defined as “a disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk.”]_{Rheme} [Although anyone can develop osteoporosis]_{Theme}, [postmenopausal women and young females with menstrual irregularities are most commonly affected]_{Rheme}. [An estimated 20% of women more than 50 years old]_{Rheme} [have osteoporosis]_{Theme}. [Although most studies have focused on women of this age-group]_{Theme1}, [osteoporosis]_{Theme2} [is potentially more deleterious in younger women because they haven’t yet attained peak bone mass, and early bone loss therefore can affect the rest of their lives]_{Rheme}.

At this point, one may naturally ask questions such as the following:

1. What is ‘information structure’? In other words, what do we want to identify? How to separate information structure from various related properties?
2. How can these information structures be identified? Is the procedure related to *any* theory of information structure?

3. How can we say whether the identified information structures are correct with respect to our goal?

The extent of discussion responding to the first question is enormous. But the foci of attention and points of view are quite diverse. Also reflecting the complexity involved in the question, it is fair to say that there are no uniformly agreed answers to this question. In addition, looking at this question from the entire span of the Identification Problem, many proposals are not sufficiently explicit for the next two steps.

The second question has received much less attention. Although several proposals have been made, each one of them has weaknesses in the coverage and/or theoretical foundation. Finally, the third question has rarely been addressed. In order to complete the entire process of the Identification Problem, this question must be answered. In the rest of this chapter, we explore these three questions in relation to previous work.

Before proceeding, it is illuminating to briefly mention closely related work by Heine [1998] and Murata and Nagao [1998]. Their focus is identification/generation of definiteness (in English) in Japanese-English machine translation. This problem is in a sense the opposite direction of the Identification Problem. But it is a problem distinct from the Identification Problem for information structure because generation of definite marking in English requires a different set of criteria. For example, we will see that definiteness marking within an embedded clause cannot be predicted from information structure (see Subsection 2.3.3).

2.2 What is Information Structure?

This section reviews the phenomenon under discussion, observes difficulties with previous definitions of information structure, and introduces a characterization of information structure that serves as the basis for subsequent discussion. At the end, the assumptions and qualifications for the present work are described.

Phenomenon under Discussion

Let us start from some observations involving a question-answer pair. Throughout this work, the **boldface** in examples indicates a pitch accent.¹

(13) *Q*: Who did Felix praise?

*A*₁: Felix praised **Donald**.

*A*₂: # **Felix** praised Donald.

*A*₃: # Felix **praised** Donald.

While the choice (*A*₁) is appropriate as a direct response to the question, the other two preceded by ‘#’ are not. The symbol ‘#’ is used as contextual inappropriateness throughout the present work, cf. the use of ‘*’ for ungrammaticality. In this case, placement of pitch accent is relevant to the delivery of information. Similarly, the following distinction can also be observed.

(14) *Q*: Who did Felix praise?

*A*₁: It was **Donald** whom Felix praised.

*A*₂: # It was **Felix** who praised Donald.

In the above case, syntax (in conjunction with intonation) has an effect similar to that of intonation in the previous example. All of the above responses in (13, 14) are grammatical, and presumably share the same propositional (truth-conditional) meaning. But they have distinct felicity conditions depending on the phonological or syntactic realization. This observation about a direct response to a question lets us believe that there is a pragmatic aspect in addition to truth-conditional semantics, which may be realized in distinct linguistic forms. This way of checking information structure is commonly called the **question test** [e.g., Sgall, 1975]. While the question test is useful for informal analysis of information structure, we do not adopt the position that the question test can always be used to identify information structure. There are complicated cases. For example, a response to a question may be embedded in a complex utterance, or responses to multiple questions may be combined into an utterance. We will explore a theory of information structure that captures the intuition behind the question test but also applies to arbitrarily complex structures in expository texts.

¹In many papers, a pitch accent is indicated by UPPERCASE or SMALL CAPS. When we cite examples from them, these conventions are translated into **boldface**. In this and the following examples, all occurrences of pitch accent correspond to H* tone in the notational system of Pierrehumbert and Hirschberg [1990].

The phenomena related to information structure are observed in various languages in a number of ways. In English, the function of intonation related to the above point is reported in Pierrehumbert and Hirschberg [1990, Sections 5.1 and 5.3]. Certain types of pitch accents, e.g., represented as L+H* and H*, are argued to have distinct functions related to the contrast seen in (13) [Steedman, 1991a]. In addition, various syntactic forms such as topicalization, left dislocation, cleft, VP preposing, inversion, heavy NP shift, *since/because*, etc. have been studied in this connection [Prince, 1984; Ward, 1990; Birner, 1994; Lambrecht, 1994; among others]. These and other types of syntactic realization are extensively discussed in, e.g., Lambrecht [1994]. More visible relations to syntactic structure are observed as word order in Catalan [Vallduví, 1990], Czech [Sgall et al., 1986], Hungarian [Kiss, 1987], Russian [King, 1995; Paducheva, 1996], Turkish [Hoffman, 1995, citing earlier work], Polish [Styś and Zemke, 1995], and Finnish [Vallduví and Vilkuna, 1998]. Another form of realization is through morphology in Japanese [Kuno, 1972], and Korean [Wee, 1995]. Vallduví and Engdahl [1996] present an extensive cross-linguistic review also including Dutch and German.

The above observation urges us to derive a general description of the phenomenon across languages. Since linguistic realization is quite diverse, it is reasonable to consider that such linguistic marking is arbitrary [Prince, 1998, p. 282].

Returning to an earlier example repeated below, we assume that the informational statuses of “*Felix praised*” and “*Donald*” are distinct.²

(15) *Q*: Who did Felix praise?

A: [Felix praised] [**Donald**].

And this informational contrast affects the felicity of the utterance. Although the above illustration uses a question-answer context for presentation purposes, the same phenomenon is observed in written texts, as in (12) in the previous section. Due to a lack of prosodic information in texts, languages like English lose certain properties that may be marking information structure. In some cases, punctuation may be used to supplement prosody. But other languages that mark information structure non-prosodically may retain more linguistic properties relevant to information structure. Considering that reading in English does not seem to suffer from lack of direct information-structure marking, we assume that there is an underlying mechanism of identifying information

²A related but distinct notion of information structure is developed in Roberts [1996, 1998].

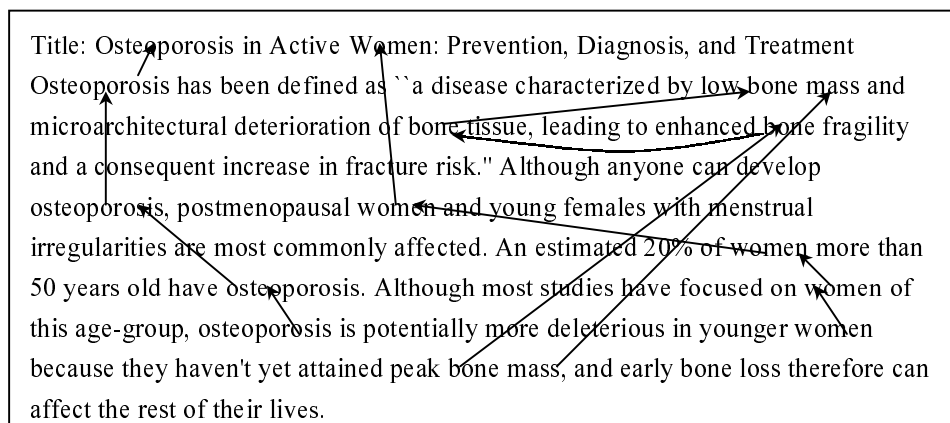


Figure 2.1: Text Link

structure that works for all the cases including written English.

At this point, we should note that our notion of information structure is orthogonal to the notion of ‘discourse topic’ [Brown and Yule, 1983, Section 3.3 (for review)]. An illuminating (informal) definition of **discourse topic** is that it is the title of a text [Brown and Yule, 1983, p. 71]. In general, discourse topic is a phrase (or a proposition, depending on the definition) associated with a text, and is *not* about the informational contrast within an utterance. As a consequence, a discourse topic may or may not be the theme of an utterance.

There is another group of work also orthogonal to the present approach. This group applies statistical methods to analyze text link (their ‘topic’) in a large corpus for speech recognition [Sekine, 1996; Jokinen and Morimoto, 1997] and discourse segmentation [Reynar, 1998]. The idea of text link is shown in Fig. 2.1. The focus of this group is a *macro* view of the discourse, and not the utterance-internal information structure we are looking at.

Difficulty with Previous Definitions

To capture the phenomenon discussed above, let us take a look at two definitions of information structure. First, Vallduví [1990, p. 18] provides the following, as a concept underlying information structure.

(16) INFORMATION PACKAGING: A small set of instructions with which the hearer is instructed by the speaker to retrieve the information carried by the sentence and enter it into her/his knowledge-store.

This definition is too broad as a starting point to work on the phenomenon of information structure. In fact, it equally applies to ‘instructions’ for speech acts. For example, it *could* be used to describe the distinction between locutionary act (reference) and illocutionary act (conventional force associated with it) [Austin, 1962].

Here is another definition from Lambrecht [1994, p. 5].

(17) INFORMATION STRUCTURE: 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 appears to contain critical elements of information structure. But it could apply to, say, presupposition projection [Gazdar, 1979]. For the investigation of the Identification Problem for information structure discussed in the previous section, both of these definitions seem to allow arbitrary instance of a theory and implementation.

Although both of the above definitions are an attempt to clarify the long-standing vagueness associated with the notion of information structure, they are not successful as a definition of information structure. To avoid problems like this, even the top-level characterization of information structure should mention the involved components and properties associated with them.

Information Structure as Semantic Partition between Theme and Rheme

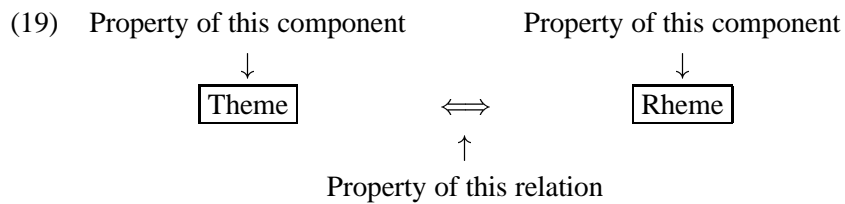
Let us first observe Vallduví’s [1990, p. 3] intuition behind information packaging: “speakers seem to structure or package the information conveyed by a sentence at a given time-point” (following earlier work of Chafe [1976] and Prince [1986]). As a simplest model, we consider a structure of two components that differ in terms of delivery of information. For example, as seen earlier, the response in a question-answer pair exhibits this point.

(18) *Q*: Who did Felix praise?

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

As in the above, we call the two components **theme** and **rheme**, following Mathesius [1975, p. 81] and Halliday [1967, p. 211]. The choice of the terminology is mainly to avoid related, but heavily overloaded terms such as ‘topic’ and ‘focus’, or ‘old’ and ‘new’ (for an extensive review of terminologies, see Vallduví [1990, Chapter 3]). But we do not follow Halliday [1967, p. 212] who states that a theme is the utterance-initial constituent. Now, our starting point is to characterize **information structure** as the abstract representation of such an organization of informational components.

To be able to provide a solution to the Identification Problem, we need to clarify the properties associated with theme, rheme, and the relation between them, schematically shown below.



This corresponds to Vallduví’s [1990, p. 23] intuition about information structure as a *relational* notion. While Vallduví departs from the binomial partition (see in Subsection 2.3.4), we pursue this binomial model in order to maintain a clear and simple notion for the relation between theme and rheme.

We now describe a preliminary version of the main hypothesis about information structure as follows:

(20) **Main Hypothesis** (preliminary version)

- a. The theme is always linked to the context (but rheme is not necessarily linked to the context).
- b. The rheme is always contrastive, in a broad sense (but theme is not necessarily contrastive).
- c. The information structure of an utterance is a complementary, semantic partition between theme and rheme.

The first point (20a) basically follows many previous proposals [Chomsky, 1971; Jackendoff, 1972; Sgall et al., 1986; Rochemont, 1986; Prince, 1992]. These proposals are discussed in detail in Section 2.3.2. A more precise characterization of the idea awaits Chapter 3.

A traditional characterization of rheme is to associate it with some kind of ‘newness’ [e.g., Jackendoff, 1972]. We will see that this position cannot be maintained (Subsection 2.3.1). Instead, the second point (20*b*) associates rheme with a general notion of ‘contrast’ such as proposed by Rooth [1985, (Alternative Semantics)]. This point is discussed in Subsection 2.3.2.

The third point (20*c*) says that theme and rheme are the only components. This also requires that a theme and a rheme combine into a proposition corresponding to the utterance in question.

Unlike the previous definitions (17) and (16), the characterization (20) at least clarifies the involved components and the properties to investigate.

In the rest of this chapter, we review previous work in relation to this informal idea. Not surprisingly, the idea is partially shared by many previous proposals. Nevertheless, we will see that every previous proposal differs from the idea in one way or another. By the end of this chapter, we will have observed that we cannot just adopt a single previous proposal as a basis for formalization and implementation along the line of (20). The main hypothesis (20) is then made more precise in the next chapters. Before moving on to the literature review, let us discuss some assumptions and qualifications.

Assumptions and Qualifications

As Vallduví [1990, Section 2.3] reviews, study of information structure is connected to various areas of linguistic studies. The course of the present work, therefore, must focus on the issues most strongly connected to the Identification Problem. We state some qualifications for the following four areas: contrastiveness, inference, reference resolution, and discourse structure.

In the main hypothesis, contrastiveness is an essential property associated with rheme. Although we review the literature in this respect, we exclude formalization and implementation of contrast. For one thing, contrastiveness is a topic on its own, which deserves a separate study [e.g., Rooth, 1985]. For another thing, its implementation is extremely difficult [Prevost, 1995 (for a small-scale implementation)]. In practice, we can achieve results useful for practical applications, as demonstrated in later chapters.

As we will see shortly in Subsection 2.3.1 (and in other sections as well), the notion of ‘contextual link’, the required property of theme, involves inference. While we discuss the way inference is involved in the Identification Problem, we exclude from discussion the *mechanism* of inference.

Although inference has been well recognized as a source of linguistic activity [e.g., Grice, 1975] and an active area in Artificial Intelligence (AI) [e.g., Russell and Norvig, 1995 (a standard text)], the state of the art is not yet at the level that we can incorporate it into our theory of information structure. Our position is that study of information structure can be done sufficiently well for practical merits without depending on the understanding of general mechanism of inference, and that the places where we fail are due to the cases where even the state of the art in the inference study does not offer a general solution.

Next, we assume that the result of reference resolution is available prior to analysis of information structure, and exclude the discussion of reference resolution itself. Reference resolution is another difficult problem on its own, theoretically and practically [e.g., Grosz et al., 1995; Hobbs, 1979]. For the purpose of identifying information structure, not knowing the correct referent does not necessarily pose a problem. For example, reference resolution of a definite expression is in general a challenging problem, but a definite expression generally provides sufficient information for the purpose of identifying information structure. That is, it in general refers to *some* entity in the context.

Finally, it is often argued that the discourse structure prior to an utterance affects reference resolution in the utterance [Grosz and Sidner, 1986; Mann and Thompson, 1988]. Now, suppose a case where multiple information structures are ambiguously available (i.e., consistent with the theory). In a way similar to reference resolution, it is quite possible that the discourse structure prior to an utterance may affect disambiguation of the available information structures. We limit our discussion to a theory of information structure that admits possible information structures, much like the way a competence grammar licenses all (and only) grammatical sentences. Although we exclude disambiguation by discourse structure, our implementation includes some heuristics for disambiguation for practical reasons.

2.3 Previous Theories of Information Structure

In his influential textbook, Levinson [1983, p. x] casts a doubt on information structure in the following manner: “the whole area may be reducible to a number of different factors: to matters of presupposition and implicature on the one hand, and to the discourse functions of utterance-initial

(and other) positions on the other.” This is a question crucial for the study of information structure, and the discussion continued until Vallduví’s [1990] demonstration against the proposition. Since this point illuminates the characteristics of information structure, this section reviews previous work mainly in relation to related properties, to which information structure was considered reducible.

The main goal of the review is to examine theories of information structure for application to the Identification Problem. Accordingly, we will pay close attention to the following check points: (1) Is the Identification Problem acknowledged? (2) Is the coverage of a theory good for realistic texts? and (3) Is the proposal under consideration sufficiently explicit for formalization and computational implementation? At the same time, this review shows that no theory singly delineates the properties addressed in the characterization (20).

In the rest of this section, we discuss information structure in relation to referential status, contrastiveness, and linguistic form. The last subsection discuss several proposals on how to partition information structure.

2.3.1 Referential Status of Theme and Rheme

In this subsection, we review the literature in the following way. Theme and rheme must be seen in relation to some referential property. But we reject the idea that information structure is reducible to referential status. After a closer look at referential status, we revisit the property of theme in connection to inference. The conclusion of the subsection is that we can capture the property of theme in relation to inference but, without depending on the problem of inference itself. At the end, we also discuss the semantic types of referents.

‘Functional’ Approaches: Recognition of Contextual Effect

The use of the terms ‘theme’ and ‘rheme’ dates back to Mathesius’s [1975, p. 81] manuscript from 1920s (Mathesius cites even earlier work), replacing more obscure terms ‘psychological subject/predicate’. The properties of theme and rheme are characterized informally as ‘given’ and ‘new’, respectively [p. 82]. Thus, by this time, properties of theme and rheme in relation to referential status had already been observed. The major contribution of the work is a clear separation of information structure from propositional (truth-conditional) meaning, and its analyses in relation

to word order (linguistic form), esp. in Czech. Mathesius calls the approach Functional Sentence Perspective (FSP) and stimulates the Prague School linguists and others to date. Halliday [1967] develops an extended system of functional (systemic) grammar. The general approach of Halliday has been applied to natural language understanding [Winograd, 1972] and generation [Matthiessen and Bateman, 1991]. Another proposal directly following FSP is due to Kay [1975], but this line has not been followed up very much. Kuno [1978] also extends this tradition and discusses pragmatic effects on English and Japanese grammar. We will come back to Kuno's work in Chapter 5.

One problem with FSP is that the properties of theme and rheme are not clearly characterized in Mathesius [1975] and also in many of the Prague school research, as mentioned in Contreras [1976, p. 16]. This tendency is still observed in more recent work including Sgall et al. [1986]. Sgall et al. [1986, Section 3.4] define 'topic' and 'focus' (corresponding to 'theme' and 'rheme') partly in terms of the notions 'Contextual Bound' (CB) and 'Non-bound' (NB) [Sgall et al., 1986, p. 178]. But the notions of CB and NB escape further clarification. They provide an operational criteria to distinguish the two that "may be found in the question test and in similar procedures" [p. 86]. Recently, an attempt of formalization has been made. For example, Peregrin [1996] describes information structure (their 'topic-focus articulation' or TFA) concisely and clearly [4. and 5. on p. 237]. This is a welcome direction, as we can evaluate the theory. But Peregrin's [1996, p. 239] formalization is too limited, as it states that the subject of an utterance (in English) is connected with a presupposition. But, as we have seen in (6) on page 5, the subject in English can be a rheme. In this regard, Halliday's [1967, p. 212] characterization of 'theme' as the utterance-initial constituent is not realistic either. We have already seen that information structure is more flexible. Another characterization of theme in Halliday [1967, p. 212] as 'point of departure' hardly delineates the involved idea.

Although theme/rheme properties in this tradition are not as clear as they should be, FSP researchers are well aware of the Identification Problem. Daneš's [1974] analysis of thematic progression, a kind of discourse structure that connects a theme to an element in the discourse, is applied to real texts. Thus, at least the theme of each utterance must be identified. The idea of thematic progression has been applied to machine translation [Papegaaïj and Schubert, 1988]. But the exposition of this material is not explicit enough for me to evaluate the effectiveness and

correctness of the procedure. Hajičová et al. [1995] along with the associated earlier work provides a computational procedure to identify information structure, to which we will return in Section 2.4.

Outside the above-mentioned work, the Identification Problem is rarely acknowledged in the theoretical studies. A common method of fixing the information structure of an utterance is to apply the ‘question test’ [Sgall, 1975]. Several proposals extend this idea and assume ‘implicit questions’ to analyze information structure in texts [e.g., van Kuppevelt, 1995, p. 110; Roberts, 1996, p. 93; Büring, 1997a, p. 178]. They hypothesize that every utterance in a text has a corresponding implicit question. The most fundamental problem with this approach is that it simply sidesteps the issue to another area. None of these analyses offers a precise way to identify the right implicit question. In addition, if we need to consider a set of ambiguous implicit questions, the set could be unbounded due to all sorts of, say, adverbial questions, unless it is constrained in a certain way. I am not aware of any practical use of this approach, e.g., text analysis or implementation.

As for coverage of realistic data, FSP researchers vary greatly. While the study of thematic progression [e.g., Daneš, 1974] commonly analyzes real data, more theoretical analysis such as Sgall et al. [1986] deal with mostly short prepared examples. In the former case, it is not clear how to identify thematic progression, and in the latter case, it is not clear whether their analysis can generally cover realistic data.

Information Structure cannot be Reduced to Referential Status

As the connection between theme and context is observed by FSP researchers, a thought was developing that information structure might be reduced to other properties [Levinson, 1983, p. x]. Chafe [1976] compared notions such as ‘givenness’, ‘contrastiveness’, ‘definiteness’, ‘subjects’, and ‘topics’. But Reinhart [1982] and von Stechow [1981] seem to give the clearest argument against information structure being reduced to referential status. Subsequently, this point is adopted by Vallduví [1990, Subsection 2.3.2] in favor of his analysis of information packaging as an autonomous level of representation.

The following example taken from Reinhart [1982, p. 18] demonstrates the point that information structure is not just referential status:³

³A similar example is found in von Stechow [1981, p. 96], which is actually a response to an earlier version of Reinhart [1982].

(21) *Q*: Who did Felix praise?

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

Reinhart [1982] points out that the referent of *Felix* and *himself* are identical. But the information structure indicated in the example is fairly clear from in this type of question-answer context. This results in a situation where both the theme and the rheme have the same referential status. Rochemont [1986, p. 52], building on Culicover and Rochemont [1983], suggests a related idea in a different way. He distinguishes two types of rheme (his ‘focus’): ‘presentational’ and ‘contrastive’. *Presentational* rheme is roughly a ‘new’ element and *contrastive* rheme is not ‘new’ (or ‘c-construable’ in his terminology) and stands in contrast to some other element.⁴ This implies that the referential status of a rheme cannot be fixed. The same point that information structure is not just reference is also made by Hoffman [1998] as she compares the roles of information structure and reference resolution applying a Centering-based theory [Grosz et al., 1995].

After separating information structure from referential status, Reinhart [1982] attempts to characterize theme in terms of the notion of ‘aboutness’ within formal semantics, adopting Stalnaker’s [1978] idea of ‘contextual set’. But such a notion is inherently knowledge-level, and requires powerful mechanism of inference, as studied in the area of Artificial Intelligence (AI). Formalization of this kind does not necessarily make the situation more explicit.

There is another attempt to provide a means of integrating information structure within semantic representation [von Stechow, 1981]. This is an important step, and we follow some of his ideas. But the discussion is limited to question-answer context and ignores the critical elements of information structure in real texts.

More on Referential Status

We have started from an intuition developed by FSP that information structure is related to referential status, but rejected the possibility that information structure *is* referential status. One important development about referential status in this connection is that there are more than just ‘old’ and ‘new’.

Prince [1981] analyzed three distinct notions of ‘givenness’ floating around at that time: (i) givenness in terms of predictability/recoverability [Halliday, 1967; Kuno, 1972], (ii) givenness in

⁴Choi [1996, p. 97] cites Dik for a similar distinction between ‘completive’ and ‘contrastive’ foci.

terms of saliency [Chafe, 1976], and (iii) givenness in terms of ‘shared knowledge’ [Clark and Haviland, 1977]. After noting the subsumption relation between these three rather heterogeneous notions, she proposed a taxonomy in terms of ‘assumed familiarity’, distinguishing EVOKED, INFERRABLE, and NEW referents. Note that we use SMALL CAPS for these terms throughout this thesis to identify the usage as we are discussing here. EVOKED referents are those textually or situationally evoked in the discourse. INFERRABLE referents are those not evoked in the discourse but the speaker believes that the hearer can infer through non-linguistic means, such as world knowledge. Finally, NEW referents are those new to the hearer (BRAND-NEW) or those known by the hearer but neither evoked in the discourse nor inferred (UNUSED). Among these three types, it is inferrable that complicates the situation most, due to involvement of inference.

Prince [1992] also introduces the notion of **discourse status: discourse-old** vs. **discourse-new** depending on whether the referent is introduced in the discourse. Yet another notion is **hearer status: hearer-old** vs. **hearer-new** with respect to the speaker’s belief about hearer’s knowledge. The terminology introduced above is summarized in Table 2.1.

Class	Subclass	Discourse status	Hearer status
EVOKED	Textually EVOKED	Old	Old
	Situationally EVOKED	New	Old
INFERRABLE		New	Old/New
NEW	UNUSED	New	Old
	BRAND-NEW	New	New

Table 2.1: Taxonomy of Assumed Familiarity (adapted from Prince [1981, 1992])

The notion of inferrable is closely related to ‘bridging’ [e.g., Clark and Haviland, 1977], and is also captured by more general notions of ‘accommodation’ [Lewis, 1979] and ‘presupposition’ [Beaver, 1997, for an extensive review].

Revisiting the Referential Status of Theme

The earlier discussion shows that the referential status of rheme cannot be fixed. But, now that we know more about referential status as seen above, we should be able to say more about theme.

Reinhart [1982, p. 21] separates theme from ‘oldness’ by excluding INFERRABLE (her ‘semantic link’) from her ‘old’. But INFERRABLE and EVOKED referents typically share linguistic

marking such as definite expression for NPs [Heim, 1982]. It is also argued that for a NP, inference is invoked by definite expression when the referent is not readily available [Bos et al., 1995; Poesio and Vieira, 1998]. Birner [1997] argues that VPs and adverb phrases too share linguistic marking between EVOKED and INFERRABLE. Considering these cases, it seems more problematic to completely separate INFERRABLES from EVOKED.

Following Reinhart, Vallduví [1990, p. 25] also separates themehood from discourse-oldness. He states that information packaging is *orthogonal* to referential status [Vallduví, 1990, p. 26]. But we need to take a closer look at this point. Vallduví [1990, p. 26] himself discusses that hearer-oldness as a *necessary* (but not sufficient) condition for topichood. Then, neither of them are in fact against the idea that theme is *not* BRAND-NEW, i.e., some combination of EVOKED and INFERRABLE.

Let us consider EVOKED and INFERRABLE themes in the following two examples:

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

ii. [The house]_{Theme} [looks exotic]_{Rheme}. (EVOKED)

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

ii. [The door]_{Theme} [looks exotic]_{Rheme}. (INFERRABLE)

For both of the above responses, it is natural to identify analogous information structures.

This observation is consistent with many other characterizations of theme/rheme (and related) distinctions. For example, Chomsky [1971, p. 199], Jackendoff [1972, p. 230], and Zubizarreta [1998, p. 1] discuss ‘presupposition’ roughly corresponding to our theme, but is distinct from the usual notion discussed by Beaver [1997]. Their ‘focus’ corresponds to our rheme in that it is informationally in contrast with theme (their ‘presupposition’). But they explicitly state that ‘focus’ is “the information in the sentence that is assumed by the speaker not to be shared by him and the hearer [Jackendoff, 1972, p. 230] and “nonpresupposed part of the sentence” [Zubizarreta, 1998, p. 1]. This distinction is basically the one between EVOKED/INFERRABLE vs. BRAND-NEW. Note that we have already rejected the simplistic characterization of rheme as BRAND-NEW [cf., Jackendoff, 1972]. Sgall et al. [1986, p. 178] distinguish ‘Contextual Bound’ and ‘Non-Bound’ (page 25). Although they do not give a precise definition, Contextual Bound seems to share the property of EVOKED/INFERRABLE. Rochemont [1986, p. 47] introduces the notion of ‘c-construable’, which again appears to be very close to EVOKED/INFERRABLE. To some extent,

this also corresponds to hearer-old [Prince, 1992, Section 2.2.2] and the idea of ‘shared topicality’ Gundel [1985].⁵

Then, we should not completely abandon the relation between information structure and referential status as Vallduví [1990, p. 25] states, but should take advantage of the relation between theme and EVOKED/INFERRABLE observed by many researchers. The tentative conclusion here is that the property of theme we mentioned in (20) is related to the referential status EVOKED/INFERRABLE.

Difficulty with Inference

If INFERRABLE is involved in the property of theme, we need to address the issues involving inference. Naturally, this is a difficult task, as can be seen in a few proposals discussed below.

Reinhart [1982, Section 6.4] observes the role of INFERRABLE (her ‘semantic link’), but does not explicate how to deal with it. Rochemont [1986, (30), p. 47] starts his definition of ‘c-construability’ in a fairly formal manner: “A string P is *c-construable* in a discourse δ if P has a semantic antecedent in δ .” Then, another definition for ‘semantic antecedent’ (31): “A string P has a semantic antecedent in a discourse δ , $\delta = \{\phi_1, \dots, \phi_n\}$, if, and only if, there is a prior and readily available string P' in δ , such that the uttering of P' either formally or informally entails the mention of P .” But, then, formal/informal entailment does not get the same level of explicitness.

Bos et al. [1995] analyze the problem of reference within the framework of Discourse Representation Theory (DRT) [Kamp, 1981]. Bos et al. [1995, Section 3.3] classify three kinds of anaphoric relations:

- (24) *a.* An antecedent is available in the discourse
- b.* An ‘implicit’ antecedent is available in the discourse (after failing the previous step): bridging⁶
- c.* No antecedent is available in the discourse (after failing the previous step): accommodation

Integrating a constrained form of inference this way has limitations. According to Bos et al. [1995], the shift from (*b*) to (*c*) depends on the availability of a *suitable* anchoring referent. But the

⁵Additional references related to this point include: Dryer [1996], van Kuppevelt [1996]. But we do not consider hierarchy of activation levels, cf. Chafe [1994].

⁶Jäger [1996] has a formal account of bridging based on dynamic semantics.

inference process involved in bridging is presumably a *general* logical process. Then, how can a system know when to fail? On the other hand, while their accommodation always saves the reference in question, we know that accommodation *can* fail. It seems more reasonable to assume that bridging and accommodation are not that different as proposed by Bos et al.

The conclusion of this subsection is as follows. Although information structure cannot be reduced to referential status, theme still has a property that is based on referential status involving inference. The previous work reviewed here fails to explicate this observation and thus not applicable to the Identification Problem. What we need is a theoretically sound, yet formalizable/implementable idea for this condition.

2.3.2 Information Structure vs. Contrast

In this subsection, we separate the notion of contrast from rheme and characterize rheme in terms of a general notion of Alternative Semantics [Rooth, 1985] that can be applied to both contrast and rhemehood.

Distinct Notions Associated with ‘Focus’

The term ‘focus’ is heavily overloaded. Thus, it is important to delineate various notions associated with it. ‘Focus’ as used by Sgall et al. [1986] and Vallduví [1990] basically corresponds to our ‘rheme’. Another group of researchers [e.g., Ladd, 1996, p. 160] use ‘focus’ as a notion closely linked to phonological properties readily observed at the word level, independent of information structure. While we distinguish these two notions, a more important point is actually to relate these two notions in a systematic way. Note that so-called ‘AI-focus’ [Grosz and Sidner, 1986, p. 179] is a way to organize referents based on their salience and should be considered distinct from other uses of ‘focus’ [Vallduví, 1990, p. 46].

The intuition we start from is that information structure is about the informational relation between units within an utterance and contrastiveness is a relation about referents not limited to those within an utterance. Thus, a rheme must always be seen in relation to a theme (possibly null) and a contrast must always be seen in relation to another referent in the context (see Fig. 2.2).

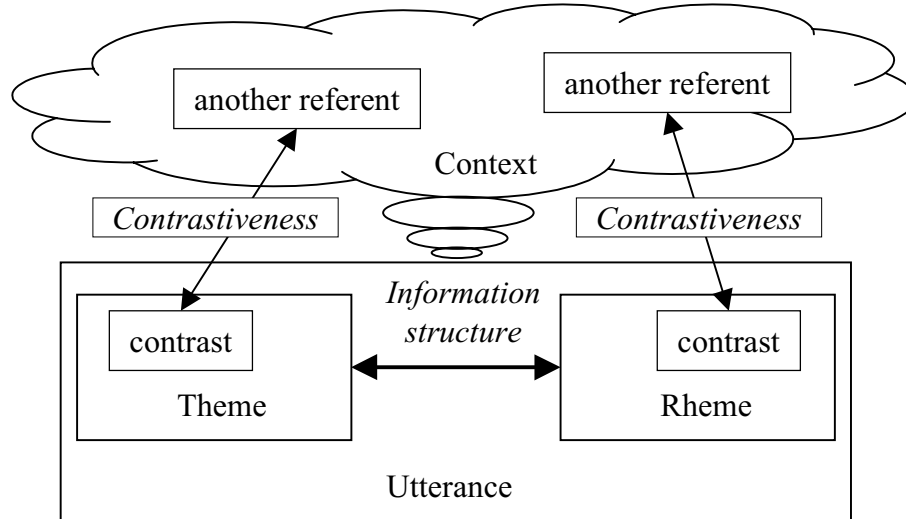


Figure 2.2: Information Structure vs. Contrast

Contrast in Relation to Phonological Prominence

First, let us explore the notion of ‘contrast’ in relation to **phonological prominence**. We consider a phonological notion of prominence at the perceptual level involving pitch, loudness, duration, and quality [Laver, 1994, p. 450], e.g., in relation to pitch accent (in English) [Ladd, 1996, p. 46, citing Bolinger (1958) and Pierrehumbert (1980)]. We continue to use **boldface** to indicate a word (in examples) where a prominence falls, as in the following example [Ladd, 1996, (5.1), p. 162].

(25) I didn’t give him three francs, I gave him **five** francs.

We use the term prominence, rather than pitch accent, to cover cross-linguistic variation in realizing the similar notion in potentially-distinct acoustic properties. Then, as in Ladd [1996, p. 160], “[i]t is now generally accepted that sentence accentuation reflects – in some way – the intended *focus* of an utterance”. This position is also taken by Jackendoff [1972, p. 229] and Gardent and Kohlhase [1996]. In the present work, we use **contrast** (instead of focus) for the semantic effect associated with prominence *on a word*. In the above example, the prominent word *five* is in contrast to *three*. On the other hand, we may call the complement of a contrast **background**.

Projection of Contrast

The notion of contrast is complicated because it can ‘project’ to a more complex linguistic structure.⁷ For example, Ladd [1996, p. 162] distinguishes between ‘narrow focus’ corresponding to our contrast on a phonological word and ‘broad focus’ spanning a more complex structure such as “*five francs*” in the following example [Ladd, 1996, (5.4)]:

(26) I didn’t give him a sandwich, I gave him five **francs**.

Even for the case where only *francs* is prominent, the phrase “*five francs*” (its interpretation) is in contrast with “*a sandwich*” in this example. As stated in Ladd [1996, p. 161], this is a phenomenon distinct from word-level ‘contrast’.⁸ Accordingly we may specifically distinguish **projected contrast** from (word-level) contrast.

As Krifka [1992] points out, and Halliday [1967] and Steedman [1991a] state more explicitly, there appears to be a connection between information structure and contrast. One complication arising from contrast projection is that a rheme may coincide with a broad focus or a single-word contrast. This is the intersection of (possibly projected) contrast and information structure.

Contrast within Theme and Rheme: Two-level Analysis

We now demonstrate that the notion of contrast (at the word level) needs to be considered independent of information structure.

Steedman [1999, (31)] provides the following example involving separation of the two notions.

(27) *Q*: I know that Mary envies the man who wrote the musical.

But who does she **admire**?

A: [Mary **admires**]_{L+H*}*Theme* [the woman who **directed** the musical]_{H*}*Rheme*.

Note: ‘L+H*’ and ‘H*’ are argued to mark theme and rheme, respectively [Steedman, 1991a].

⁷Winkler [1997] is a good review on focus projection especially in connection to syntactic structure. Another recent work is Gussenhoven [1999].

⁸Hockey [1998, p. 226] discusses the role of amplitude and duration in marking the entire span of rheme (her ‘focus’) in English and Hungarian.

In addition to the projection of contrast from *directed* in the rheme, there is another contrast *ad-mires* in the theme.⁹ The two instances of contrast above receive distinct pitch accents corresponding to theme and rheme [Steedman, 1991a]. Halliday [1967] too discusses the two levels: ‘information structure’ [p. 199] roughly corresponding to our information structure and the distinction between ‘new’ and ‘given’ [p. 204] (corresponding to our contrast/background).¹⁰

Vallduví and Vilkuna [1998, p. 85] also distinguish ‘rheme’ (‘focus’ in Vallduví [1990]) from ‘kontrast’ (their new terminology). But their kontrast is a notion associated with a constituent, and thus is intermediate between our rheme and our (word-level) contrast. Their analysis would make the projection problem of contrast unnecessarily complicated.

The distinction between rheme and contrast is not always understood as the above. Since the term ‘focus’ is overloaded, analyses often mix the two notions. For example, Pulman [1997, p. 74] uses the term ‘focus’ citing the Prague School (some work preceding Sgall et al. [1986]), Selkirk [1984], and Rooth [1985]. But this introduces a complication because ‘focus’ of the Prague School basically corresponds to our rheme, and that of Selkirk [1984] and Rooth [1985] corresponds to our contrast (and its projection). The subsequent description of broad and narrow foci does not illuminate the discussion. He distinguishes narrow and broad foci based on constituent size, which is misleading. It is not clear why his approach works for the case of broad focus without discussing focus projection.

Property of Rheme: Projection from a Contrast

We have shown above that a theme *can* contain a contrast. But, as seen in many earlier examples, a theme does not always contain a contrast. But a rheme is always projected from a contrast [e.g., Jackendoff, 1972, p. 229; Rochemont, 1998, p. 337]. The main point in this subsection is to examine (20c) of the main hypothesis: “a rheme is always contrastive”. At this point, let us recall Rochemont’s [1986, p. 52] distinction between ‘presentational’ and ‘contrastive’ rhemes. This suggests that there is a rheme that is *not* contrastive. For example, in the example (18) repeated below, the response may be considered to include a presentational rheme (without further

⁹Prevost [1995, p. 67] calls the contrasts in a theme and a rheme ‘theme-focus’ and ‘rheme-focus’, respectively.

¹⁰Fries [1994, p. 234] calls Halliday’s information structure and given/new distinction thematic structure and ‘information structure’, respectively. Brown and Yule [1983, Chapters 4 and 5] use theme/rheme in the sense of Halliday [1967], and information structure for Halliday’s [1967] given/new.

contrasting information).

(28) *Q*: Who did Felix praise?

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

But Jackendoff [1972, p. 246] observes that a negative response such as the following is possible in the same context.

(29) *Q*: Who did Felix praise?

A: [Felix praised]_{Theme} [**nobody**]_{Rheme}.

This suggests that there is no presupposition for the existence of an individual who was praised by Felix. The rheme in (28), *Donald*, is in contrast at least with *nobody*. Thus, as argued by Büring [1997b, p. 40], it is possible to abstract away from Rochemont's distinction.

Alternative Semantics

There is a general way to capture the semantics of contrast, i.e., Alternative Semantics [Rooth, 1985]. The idea is that the notion of contrast can be defined by considering an 'alternatives' set where the elements in contrast are marked. For example, for an alternatives set {"[John]_c is tall", "[Mary]_c is tall"}, *John* is in contrast with *Mary*. In other words, the alternatives set is obtained by making an appropriate substitution in the contrastive element. The selection of these contrastive elements can span an arbitrarily long distance across utterances in a discourse (or in the context in general). Therefore, the exact nature of how such contrast is analyzed is obviously beyond the grammar for the utterance level [discussion in Rooth, 1992]. Partee [1999, p. 214] comments on this point that Rooth [1992] is an extreme of degrammaticalized analysis of contrast (her 'focus'). But, since contrast spans across discourse (and situational context as well), it is natural and necessary for a theory of contrast to have a degrammaticalized component.

An advantage of Alternative Semantics is that it can be applied to a projection of contrast in a general way. Now that we consider a rheme as a projection of contrast, the rheme can be seen in terms of the alternatives set associated with the theme [Steedman, 1999, Section 5.3]. It is this view that a property of rheme is contrastiveness in a general sense (20c).

But Alternative Semantics does not automatically solve the problem of identifying rheme through contrastiveness. It is a general framework that can be used for accounting for the semantics

of (narrow and projected) contrast *and* that of rheme. To complete the analysis of contrastiveness, we must have a mechanism of identifying the alternatives set, which is extremely difficult to formalize and implement. On the other hand, in order to apply it to the Identification Problem of information structure, we also need to know the relation between rheme and theme. This latter point is not clearly stated in Rooth [1992, p. 84] when he argues that Alternative Semantics can be applied to the analysis of question-answer context. We will address the relation between theme and rheme in the next chapter.

Dynamic Semantics: Connection to Procedural Accounts

Alternative Semantics can also be connected to procedural ideas through ‘dynamic semantics’ [Stalnaker, 1978, (an earlier work)]. In this tradition, the meaning of an utterance is considered as a potential to change context. The representation of context differs among proposals. For example, Stalnaker [1978, p. 321] has it as a set of propositions. Heim [1982] has it in terms of files in File Change Semantics (FCS). Kamp [1981] has it as Discourse Representation Theory (DRT). More recent work relevant to our case are: Asher [1993] and Hendriks and Dekker [1996].¹¹ The idea of dynamic semantics is adopted in recent analyses of information structure McNally [1998, Section 3.2] and Steedman [1999, Subsection 5.3.1].

In this subsection, we have separated the notion of contrast from information structure, and observed a requirement that a rheme (semantic unit roughly corresponding to a constituent) be projected from a contrast (word-level property). While this identifies a property of rheme useful for theoretical analysis, its formalization and implementation for the Identification Problem remains open.

2.3.3 Information Structure and Linguistic Form

This subsection explores direct linguistic marking of information structure and argues that information structure cannot be obtained from linguistic form alone.

¹¹Atlas [1991] may also be included in this group.

Linguistic Marking of Information Structure as a Matrix-level Phenomenon

It is generally accepted that linguistic marking of information structure exists [e.g., Vallduví and Engdahl, 1996]. But very little has been said about properties generalizing various distinct forms of information-structure marking. One reason may be that linguistic marking of information structure is arbitrary [Prince, 1998, p. 282]. As a tool to analyze linguistic marking of information structure, I would like to examine the following hypothesis:

- (30) (hypothesis) Linguistic marking of information structure is a matrix-level ('root') phenomenon, i.e., *non-recursive*.

Naturally, this view is consistent with most proposals of information structure including our main hypothesis (20), which is non-recursive (the idea of recursive information structure is reviewed in Subsection 2.3.4). This is in contrast with the use of, say, definite determiner, which is recursive along linguistic structure. A consequence of the above hypothesis is that Levinson's [1983] complaint about lack of projection analysis for information structure is not actually applicable to information structure itself. But it may apply indirectly to information structure through other types of linguistic marking, e.g., definiteness. Let us now examine some examples of information-structure marking discussed in the literature.

First, it is generally held that prosodic structure is non-recursive [Selkirk, 1984; Pierrehumbert and Beckman, 1988; as reviewed by Ladd, 1996, p. 238]. If certain pitch accents, e.g., L+H* and H* as shown in Pierrehumbert and Hirschberg [1990] are associated with theme and rheme [Steedman, 1991a], respectively, such a pitch accent, associated with a word, may recursively project through linguistic structure. But the prosodic units projected from pitch accents do not embed another unit, as formally shown in Steedman [1999, Section 5.6]. Thus, there is no conflict between prosodic structure that marks information structure and the hypothesis (30). Note that Ladd [1996, p. 245] himself argues for recursive prosodic structure, but this means that prosodic structure can recursively associate with linguistic structure and is not a position contrary to what has been said above.

Although English does not have an extensive set of direct information-structure markers (compared to languages like Catalan), there are many special constructions whose functions have been discussed in relation to information structure. Among these, inversion cannot be embedded while subordinators *since/because* can (examples and more details in Subsection 3.3.2). The hypothesis

predicts that the former can be but the latter is not a direct information-structure marker.

A strong support for non-recursiveness of information-structure marking comes from particle use in Japanese. While the detailed discussion awaits Chapter 5, it is illustrative to point out that *thematic* function of particle *wa* is only available at the matrix level. In addition, a constituent extracted from an embedded level can also be marked in this way. Direct information-structure marking is a basis for our evaluation method. Later, we use particle choice in Japanese in the evaluation process.

Discussion of languages other than English and Japanese is beyond the scope of the present work, but I am very much interested in analyses for or against the hypothesis (30).¹² The prediction of the hypothesis is that recursive linguistic marking is not a direct information-structure marking. For example, is it really the case that a clause (IP), regardless of matrix or embedded level, is ‘configured’ according to information structure, e.g., in Russian [King, 1995]?

The theme-first principle is certainly a controversial one as a linguistic marking of information structure [Lambrecht, 1994, Section 4.5, for a detailed discussion]. There are some experimental results showing that passivization is associated with information-structure effect [Most and Saltz, 1979]. But the current work is not committed to accept that theme-first principle applies universally, or even language-specifically, as information-structure marking. But we do consider a certain cases of preposing, e.g., utterance-initial modifier, as a contextual-link marker based on de Swart’s [1999] analysis. More detail is described in Subsection 3.3.1.

Information Structure cannot be Recovered Solely from Linguistic Form

We have seen above that information structure may be marked linguistically. In this connection, Vallduví [1990, p. 6] states that “[i]nformational understanding and the packaging instructions that encode it must obviously be recoverable from the overt structure of any language”. This is a very strong statement suggesting that the linguistic structure completely identifies the information structure. We have to disagree with this position following Brown and Yule [1983, p. 188] who state that linguistic form alone is not enough to identify information structure.

The following example from Steedman [1991a, p. 285] demonstrates that exactly the same

¹²Kiss [1995] discusses a number of languages in relation to the idea of ‘discourse configurationality’.

linguistic forms including prosody may have distinct information structures depending on the context.

- (31) a. [They are]_{Theme} [a good source of **vitamins**]_{Rheme}. (in response to “What are legumes?”)
 b. [They are a good source of]_{Theme} [**vitamins**]_{Rheme}.
 (in response to “What are legumes a good source of?”)

Similarly, in Japanese, exactly the same utterance including phonological marking can be ambiguous with respect to information structure (assuming that there is no phonologically marked distinction between theme and rheme, and particle *wa* can be used for a theme and contrastiveness, as will be discussed in Chapter 5). Here, the following grammatical labels are used TOPic, CONTRastive, ACCusative.

(32) Q: “What did Ken and Naomi do?”

A: [**Ken-wa**]_{Theme} [**banana-o** tabeta]_{Rheme}.
 Ken-TOP/CONT banana-ACC ate
 “Ken (but not Naomi) ate a/the banana.”

(33) Q: “Between Ken and Naomi, who ate the banana and the mango?”

A: [**Ken-wa**]_{Rheme} [**banana-o** tabeta]_{Theme}.
 Ken-CONT banana-ACC ate
 “Ken (but not Naomi) ate the banana.”

Vallduví’s [1990] position indeed suggests that information structure *cannot* be affected by the context. This reduces identification of information structure to parsing. Possibly for this reason, Vallduví [1990] does not address the problem of identifying information structure in texts, and only works on examples that do not show the problem of ambiguous information structure. Nevertheless, Vallduví and Engdahl [1994, p. 531] state that “no syntactic constituency is required for any informational unit as long as inheritance of INFO-STRUCT values proceeds in the permitted fashion”. This seems to discount Vallduví’s [1990] position that information structure can be completely derived from surface structure.

We have seen that information structure cannot be identified from linguistic form alone. We have also noted that linguistic marking of information structure is relatively impoverished in written English. But it seems that linguistic communication in written English does not suffer from

potentially ‘defective’ information structure. In the next chapter, we develop the main hypothesis (20) in terms of properties including definiteness, which is systematically employed in English.

2.3.4 Internal Organization of Information Structure

This subsection examines different ways of organizing components of information structure: i.e., recursive structure, binomial and trinomial partition, and graded multiple partitions.

Recursive Information Structure

Our main idea about information structure (20) assumes that it is non-recursive. We have also stated a hypothesis, (30), that linguistic marking for information structure is matrix-level. But some argue that information structure is recursive [i.e., Kiss, 1987; Hoffman, 1995, p. 145; Partee, 1996, p. 77].

Let us examine the following example from Partee [1996, (31), p. 82]:

- (34) What convinced Susan that our arrest was caused by **Harry** was [_{FOC1}a rumor that [_{S3} someone had [_{FOC3} witnessed Harry’s confession.]]]

Partee analyzes the structure for this utterance in the following way:

- (35)
$$\begin{array}{ccc} \text{TOP2} & \text{FOC2} & \text{TOP3} & \text{FOC3} \\ \hline & S2 & & S3 \\ \text{TOP1} & & \text{FOC1} & \\ \hline & S1 & & \end{array}$$

Partee [1996, p. 67] is specific about her ‘topic’ and ‘focus’ are Praguian [Mathesius, 1975; Sgall et al., 1986, etc.]. But there are two points we may argue against recursive information structure. First, there is no standard way to identify information structure recursively, cf. ‘question test’ [Sgall et al., 1986], which is non-recursive. Second, commonly observed direct information-structure marking is non-recursive, as we have seen for the hypothesis (30). With a focus on the contextual status of a clause, Partee’s [1996] analysis is more in line with formal analyses of ‘presupposition’ [e.g., Beaver, 1997]. The problem of presupposition projection is widely discussed in relation to linguistic structure [e.g., Gazdar, 1979; Karttunen and Peters, 1979]. Once contrastive elements [Rooth, 1985] involved in the utterance are identified, two-level analysis (page 2.3.2) of

Steedman [1991a] seems sufficient for the above example. A convincing demonstration of recursive information structure would identify a test comparable to question-test for arbitrary linguistic structure or find recursive linguistic marking that directly marks information structure.

One motivation often found behind recursive information structure is to identify information structure with tripartite quantification structure [Partee, 1996] (also to some extent in Partee [1999]). A quantification structure has the form *Quantifier (Restrictor, Scope)* commonly used in formal semantics. Applying this connection, Szabolcsi [1983b], Rooth [1985], and Sgall et al. [1986] argue that information structure is truth-conditional.¹³ Szabolcsi [1983b, Section 3.1] explicitly states exhaustivity as the cause of this point, and the same situation is implicit in Sgall et al. [1986, p. 62] as well. For this matter, I follow Horn [1981, p. 132] and Vallduví [1990, Section 7.1] in that exhaustivity is conversational implicature [Grice, 1975] (for English, not a direct counterexample to Hungarian examples in Szabolcsi [1983b]). Kuno [1972] also states the exhaustivity effect for a Japanese particle *ga*, but rejected by Shibatani [1990, p. 271] as epiphenomenal (more discussion in Chapter 5).

Binomial Partition

The rest of this subsection reviews some proposals on non-recursive information structure. The classic partition of information structure is the binomial one, e.g., early Prague School [Mathesius, 1975], and [Chomsky, 1971; Jackendoff, 1972; Halliday, 1967; Steedman, 1991a]. But its simplicity is also associated with some problems. For example, Vallduví [1990] argues that neither topic-comment nor focus-background can properly represent the partition commonly observed in natural data. In general, the complexity of realistic texts poses a challenge to binomial partition.

First, let us consider the following example from Vallduví [1990, (42), p. 55]:

(36) *Q*: What does John drink?

A: [John]_{Link} [drinks]_{Tail} [beer]_{Focus}.

Vallduví [1990] proposes a trinomial partition of information structure such that our theme is further divided into two subcomponents. His ‘link’ and ‘tail’ jointly correspond to our ‘theme’, and ‘focus’ to our ‘rheme’. His argument, then, is that focus-background partition would result in

¹³Relevant other papers include: Szabolcsi [1981], Szabolcsi [1983a], Erteschik-Shir [1997], and Erteschik-Shir [1998], Lee [1993], and Jäger [1999].

“focus = *beer*” and “background = *John drinks*” and topic-comment partition would result in “topic = *John*” and “comment = *drinks beer*”, and that neither of them capture the information structure properly. While the focus-background partition directly correspond to the informational division of the question, topic-comment structure (as presented by Vallduví) does not. There are two points to make here. One is about semantic types for referent, and the other is about accommodation of a theme. In the following, we discuss these points in turn.

Most studies of reference in relation to information structure deal only with (discourse) referents [Karttunen, 1976] of the individual type, corresponding to referential NPs. For example, Reinhart [1982, p. 5] limits the discussion of theme to NPs. This also applies to Vallduví [1990, Chapter 4] adopting an analogy of File Change Semantics (FCS) [Heim, 1982], and Hoffman [1996] adopting a version of Centering theory [Grosz et al., 1995]. But a question like (36Q) partitions information where subject-verb sequence is a unit of information, as observed by Steedman [1991a, p. 260]. In general, any linguistic units that are extractable or can be coordinated may well be an information-structure unit [Steedman, 1996]. In accordance to this observation, Vallduví and Vilkkuna [1998, p. 82] seem to have dropped File Change Semantics in favor of a more general extension of Discourse Representation Theory [Kamp, 1981; Heim, 1982], an extension due to Asher [1993] to deal with ‘abstract objects’. A consequence of this more general view of referent allows us to analyze “*John drinks*” in (36A) as a unit of information structure even though it is not traditionally considered a constituent. Thus, as long as we have a means to account for such constituents, e.g., Combinatory Categorical Grammar [Ades and Steedman, 1982], this type of division is not a problem for binomial partition. Then, we need some other explanation for separating *John* in (36A) as Vallduví’s [1990] ‘link’.

The other point is the possibility of accommodating a theme. Although a direct response to a question such as (36A) is what we usually expect, we may also encounter unexpected responses, including completely irrelevant ones. Note that question test as a tool to identify the information structure of a response is only good for a direct response. But even for non-direct response, we will find a certain information structure depending on the context. Let us consider the following example with ambiguous information structure.

(37) *Q*: Who did Felix praise?

*A*₁: [Felix praised]_{Theme} [Donald]_{Rheme}. (direct response)

A_2 : [Felix] $_{Theme}$ [praised Donald] $_{Rheme}$.

A_3 : [Felix praised Donald] $_{Rheme}$.

As long as the theme is linked to the context (including the null case) and the complementary rheme is a projected contrast, any of the above information structures are possible, which is consistent with our main hypothesis (20). Although the contextual force of a question is very strong, it cannot completely specify the response. There is a room for the respondent to *accommodate* a distinct theme (see Subsection 2.3.1 for accommodation). Thus, theoretically, the following ambiguity for (36) is possible.

(38) Q : What does John drink?

A_1 : [John drinks] $_{Theme}$ [beer] $_{Rheme}$.

A_2 : [John] $_{Theme}$ [drinks beer] $_{Rheme}$.

Note that the above analysis observes an ambiguity, but *not* a coexisting parallel structures, as in Vallduví [1990]. Without additional contextual information, (A_1) is the most likely response. (A_2) may still be available if, e.g., the context is specifically about John and elaborating various properties of John. In summary, we accept the possibility of information structure like (38 A_2), but it can be analyzed within the binomial partition approach.

There is another type of problem for binomial partition. Let us take a look at another example from Vallduví [1990, (56a)], assuming a question “What did the farmer do with the broccoli to the boss?”.

(39) [The farmer] $_{Link}$ [already **sent**] $_{Focus}$ [the broccoli to the boss] $_{Tail}$.

In this case, the theme (*Link + Tail*) is *discontiguous*. Related examples are found in Büring [1997b, (4,5), p. 3].

(40) *i.* Guess who went to the central station after Smith left the pub.

ii. After Smith left the pub, [**Jones**] $_F$ went to the central station.

Again, the theme (i.e., the complement of *John*) is discontiguous. This case is a problem for binomial partition that assumes complete syntax-semantic parallelism. But it is still possible to construct a semantic unit covering the discontiguous themes. We will explore a principled method to link such a semantic structure with syntax in the next chapter.

Trinomial Partition

In an attempt to avoid the problem with binomial partition, Vallduví [1990], Büring [1997b], and Hoffman [1995] adopt a trinomial partition.¹⁴ For Vallduví [1990] and Büring [1997b, p. 54], it is a way to mediate both topic-comment and background-focus partitions, also suggested by Jacobs [1986, p. 104].

Vallduví [1990, p. 57] proposes a trinomial partition of information structure “*Link – Focus – Tail*”. This corresponds to our “*Theme – Rheme – Theme*” case as Vallduví’s [1990] ‘link’ and ‘tail’ are in contrast with his ‘focus’, e.g. (36). But this partition does not generalize to cases such as the following [p.c., Steedman 1998]:

(41) *Q*: I know what team Fred wants to win the Cup, but which team does Alice want to lose which contest?

A: [Alice wants]_{Theme} [**Australia**]_{Rheme} [to lose]_{Theme} [the **Ashes**]_{Rheme}.¹⁵

Hoffman [1995, Chapter 5] proposes a slightly different trinomial partition “*Topic – Focus – Ground*”. But she combines ‘focus’ and ‘ground’ as ‘comment’ in contrast to her ‘topic’, and only considers contiguous partitions between ‘topic’ and ‘comment’. Thus, it is not a solution to the problem of discontinuous information structure. Similarly, Fries [1994, p. 234] divides rheme into N-Rheme (last constituent) and the rest (assuming Halliday’s theme).

We have separated the notion of ‘contrast’ from information structure, and have accepted that contrast can appear freely within a theme or a rheme [Halliday, 1967; Steedman, 1991a]. Thus, partitions between ‘contrast’ and ‘background’ within a theme or a rheme can be accounted for without problem. This approach can cover Hoffman’s [1995] and Fries’s [1994] analyses more generally. In summary, trinomial approaches do not seem to be a solution to the discontinuity problem.

Another question about these trinomial partitions is how can we *define* such further divisions of theme and rheme. It is not entirely clear how the two theme components in the above examples are distinct in a systematic manner. Although Catalan seems to split a theme across the rheme, such a distinction between the two theme components seems language-specific and does not show up in other languages in a systematic way. Hendriks and Dekker [1996, p. 350] also argue against the

¹⁴A similar observation is made in Foley [1994, p. 1680], which is a fairly extensive encyclopedia entry.

¹⁵With or without L+H* on the themes.

status of ‘tail’ [Vallduví, 1990] that it complicates analysis and processing of information structure (they show an example to demonstrate such a complication).

Communicative Dynamism

Another, more complicated approach is Communicative Dynamism (CD) [Firbas, 1964], developed within the Functional Sentence Perspective (FSP) approach. CD is a degree of contribution to the development of the communication by sentence elements. Firbas [1964, p. 272] states that Communicative Dynamism is not dependent on ‘unknown’ vs. ‘known’. Communicative Dynamism is by definition a *gradable* concept. While it may well be the case that information ordering is graded, it is hard to grasp the idea cross-linguistically in terms of observable phenomena. While information ordering may be faithfully realized in a language like Czech, it is not readily observable in other languages to the level we can generally see for the contrast between theme and rheme. Second, there is no generally accepted ‘semantics’ for such grading. Finally, in relation to the first two points, it is extremely hard to evaluate. Thus, CD is not appropriate for the current purpose. Note that we do not deny the possibility of multiple divisions. There may be factors that are beyond the current scope and have not been clarified in the previous work.

Summary

In any of the reviewed cases, there are some kinds of problems. Since additional complexities associated with multiple partitions do not solve the problem as a whole, we assume the classic and simplest case, binomial partition (Ockham’s razor). The problem with binomial partition, namely discontinuous information structure is addressed in detail in the next chapter.

2.4 Previous Proposals for Identifying Information Structure

There are several proposals directly addressing the Identification Problem [Kurohashi and Nagao, 1994; Hahn, 1995; Hajičová et al., 1995; Styš and Zemke, 1995; Hoffman, 1996; Komagata, 1998a]. This subsection reviews these proposals. We also discuss application of information structure to natural language generation at the end because this computational application too involves the Identification Problem

While each one of these approaches has particular problems of its own, there are more fundamental problems shared by these approaches: namely, limited coverage, lack of evaluation, and unclear theory-procedure relation. The following review pays close attention to these points.

Kurohashi and Nagao, 1994

The main point of Kurohashi and Nagao [1994] is that ‘discourse structure’ in Japanese in the sense of Grosz and Sidner [1986] and Mann and Thompson [1988] can be identified through surface information. Discussion of their main goal is naturally beyond our scope, but we must investigate the component involving the notion of information structure, namely the problem of identifying information structure (their ‘topic’/‘non-topic’) in Japanese. Their method basically consists of observing the distribution of particles *wa* (so-called ‘topic marker’) and *ga* (nominative marker) without using contextual information.¹⁶ Analysis and the use of these particles are important aspects of text analysis in Japanese, and we follow this direction. But the functions of these particles are complex and we cannot simply say that *wa* and *ga* mark theme and rheme, respectively (see Chapter 5 for more detail). Moreover, there are utterances lacking these particles (as arguments can be dropped in Japanese), still with clear information structure depending on the context.¹⁷ Kurohashi and Nagao [1994] are also limited in explicating the theory-procedure relation with respect to the description of (partial) relation between Japanese particles and information structure. Finally, their analysis only contains a language-specific element of information structure. Since

¹⁶Kurohashi and Nagao [1994] also apply a few additional structural cues, which are not clear from the paper.

¹⁷The following example demonstrates that information structure is not necessarily marked by *wa* or *ga* (grammatical labels: TOPic, ACCusative, and Question):

- (1) *Q*: Ken-wa Montana-to Oregon-de nani-o sita-no?
 Ken-TOP Montana-and Oregon-at what-ACC did-Q
 “What did Ken do in Montana and Oregon?”
- A*: [**Montana**-de]_{Theme(contrastive)} [**sukii**-o site,...]_{Rheme}
 Montana-at ski-ACC did
 “He **skied**_{H*} in **Montana**_{L+H*,...}.”
- (2) *Q*: Ken-wa doko-de sukii-to sukeeto-o sita-no?
 Ken-TOP where-at ski-and skate-ACC did-Q
 “Where did Ken ski?”
- A*: [**Montana**-de]_{Rheme} [**sukii**-o sita.]_{Theme}
 Montana-at ski-ACC did
 “He skied in **Montana**_{H*}.”

our position is that the notion of information structure applies cross-linguistically and that it contains universal elements, the approach of Kurohashi and Nagao [1994] does not apply to analysis of other languages. Since their goal is identification of discourse structure, no direct assessment of the information structure is provided.

Hajičová and others, 1995

Following the tradition of the Prague school, e.g., Sgall et al. [1986], Hajičová et al. [1995] proposed an algorithm to identify information structure (their ‘topic’ and ‘focus’).¹⁸ Their algorithm is an implementation of a series of theoretical works, it addresses the theory-processor relation more strongly than others.

But there still remains a question about theory-processor relation. Although they discuss a contextual factor in terms of their ‘Contextual Bound’ (CB) and ‘Non-Bound’ (NB) (p. 25 in Subsection 2.3.1), their algorithm actually assigns a CB/NB status through structural analysis [p. 89-90], as seen below.

- (42) (a) After the dependency structure of the sentence has been identified by the parser, so that also the underlying dependency relations (valency positions) of the complementations (to the governing verb) are known, the verb and all the complementations are first assumed to be NB, i.e., to belong to the focus, which we denote by f.
- (b) (omitted: three conditions for the case where the verb is rightmost)
- (c) If the verb does not occupy the rightmost position, then:
 - (ca) the verb itself is understood as t [topic], if it has a very general lexical meaning (see above), or as f if its meaning is very specific, or else as ambiguous (t/f);
 - (cb) the complementations preceding the verb are denoted as t, with the exception of an indefinite subject and of a specific (i.e., neither general nor indexical; see above) Temporal complementation; either of the latter two is characterized as t/f;
 - (cc) (omitted: ten more conditions for various cases)

The condition (cb) thus predicts that a definite subject is a topic as they do in their example (3) “The neighbor met him yesterday” [p. 91]. But, as the following example shows, a definite subject

¹⁸Two closely related papers are Hajičová [1991] and Hajičová et al. [1993].

with a verb not at the rightmost position can be a rheme.

(43) *Q*: Who met him yesterday, the neighbor or the gardener?

A: The neighbor met him yesterday. Hajičová et al. [1995, (3), p. 91]

“*The neighbor*” in (43A) must be analyzed as the theme (or its part) of the utterance. As we have discussed earlier, linguistic form alone cannot fix the information structure.

We agree that certain linguistic marking such as definiteness plays an important role in identifying information structure, and we will use that property. But we cannot underestimate the contextual effect. The algorithm depends too much on structural and lexical information and has very little contextual information in it. The coverage of the algorithm is limited to simple sentences in English. They comment on the extension of the proposal to more complex constructions [p. 93]. But their algorithm [p. 89-90] is already a sequence of *seventeen* conditional statements. Even if it can be extended to more complex cases, it would be hard to see the underlying generality. Finally, no evaluation is discussed.

Hahn, 1995

Hahn [1995] argues that thematic progression [Daneš, 1974] can be formalized, be applied to real-world texts, and provide a means to view text coherence. The implementation consists of partial parsing, processing of ‘frame’ representation including relations between entities, and processing of theme/rheme according to how the theme of an utterance is connected to an antecedent in the context. The system works on realistic data taken from computer-related journals. This approach has a strength in dealing with real data, unlike many other approaches discussed here. The contextual information is well handled as well.

The problems with this approach include the following. Although Hahn [1995, p. 215] argues that full parsing is infeasible for such a task, there is a cost associated with adopting partial parsing. For example, the information obtainable from complex NPs can be misused. In addition, special constructions such as ‘cleft’ and ‘topicalization’ cannot be identified without ad-hoc treatment. A systematic analysis of sentence construction requires full parsing. Furthermore, the system appears to be limited to individual-type themes. It could not identify a theme such as “*Felix praised*” (18) seen earlier. There is little discussion about how his implementation is related to a theory of information structure. Again no evaluation method is provided.

Styś and Zemke, 1995

Styś and Zemke [1995] proposes a method to improve the quality of English-Polish machine translation. Their point is that word order in Polish depends on salience and this information can be obtained in English through linguistic analysis including Centering theory [Grosz et al., 1995, as well as much earlier work cited there]. Their approach is actually more in line with Communicative Dynamism (CD) [Firbas, 1964] because their theory adopts ‘gradation’ of salience, not binomial contrast between theme and rheme. They obtain such results by applying gradation to Centering analysis, utterance construction type, definiteness, constituent length, etc. There is no doubt that information structure is related to most, if not all, of these properties. But the use of graded salience makes evaluation of this approach extremely difficult. Accordingly, no evaluation is discussed. Furthermore, an ad-hoc weighting of these properties does not seem to be well-founded in terms of available theories of information structure. Styś and Zemke [1995] mainly deal with the transitive construction including clefted cases [Section 5 (Conclusion)], and need to extend their limited coverage for a more realistic set of data.

Hoffman, 1996

Hoffman [1996] proposes a method to improve the quality of English-Turkish machine translation through the use of information structure. The key element of the proposal is identification of information structure in English through a combination of contextual information and linguistic form, including Centering analysis [Grosz et al., 1995]. This in principle combines the strengths of Hajičová et al. [1995] and Hahn [1995]. Hoffman [1996] characterizes theme (her ‘topic’) in terms of referential preference based on a version of Centering theory [Grosz et al., 1995], and rheme (her ‘focus’), in terms of ‘discourse-newness’ and ‘contrastiveness’, corresponding to the distinction of Rochemont [1986].

The main contribution of the proposal is the following two algorithms:

(44) Topic algorithm:

a. Choose C_b (if available) as the topic.¹⁹

¹⁹C_b is the highest-ranked referent in the reference list (C_f) of the previous utterance also present by the current utterance.

- b. Choose the first entity in the Cf list (if available).²⁰
- c. Choose a situation-setting adverb (if available).
- d. Choose the subject.

(45) Focus algorithm:

- a. Choose a discourse-new
- b. Choose a contrastive element

Use of these algorithms is demonstrated in Hoffman's (5), which can be shown as follows ('topic' and 'focus' are indicated with the rule that is used to identify it):

- (46) *i.* Pat will meet Chris today .
Focus (45a) Focus (45a) Topic (44c)
- ii.* There is a talk at four .
Focus (45a) Topic (44c)
- iii.* Chris is giving the talk .
Focus (45b) Topic (44a)
- iv.* Pat cannot come .
Topic (44b) Focus (45a)

One of the weaknesses of Hoffman's algorithms is its lack of connection to a theory of information structure. For example, it is not at all clear why *today* in (i) *must* be the topic. Information structure is characterized in terms of combination of referential status and other properties on the involved components. It does not capture the *relation* between theme and rheme in a way we are interested in.

Another problem is its limitation in recognizing 'referents' corresponding to complex linguistic structures. In the following example similar to the one given in her paper, the theme algorithm will pick up "Chris" as the theme of (ii), among the possible candidates underlined below.

- (47) *i.* Chris will give the talk. [Chris, talk]
- ii.* But, Pat doesn't think that Chris will give the talk. [Pat, Chris, talk]

But the clause "that Chris will give the talk" is most likely the theme of (ii).

Hoffman [1996] tackles cases involving adverbs and complement clauses, but demonstrates her algorithm only for a few prepared texts, not realistic data. She also mentions the role of INFERRABLE, which is a critical element in identifying information structure, but does not specify how to identify them. Finally, there is no evaluation is presented.

²⁰Cf is the list of discourse referents in the utterance.

Komagata, 1998

In the precursor to the current work [Komagata, 1998a], I proposed a theory of information structure and an algorithm to identify information structure to be used for a Computer-Assisted Writing system. The goal of the system is to detect text readability with respect to information structure. The mechanism of the identification process is that theme has a property ‘contextual link’, which is realized as either discourse-old or linguistically-marked inferrables like Hoffman [1996]. Then, the theme-rheme structure is observed as the last semantic composition.

Some problems with this work are that the theory is overly simplistic. For example, the only considered linguistic marking for inferrable was definiteness. The theory assumed binomial partition of information structure where theme and rheme are contiguous, which is not necessarily the case (see Subsection 2.3.4).

In an attempt to address lack of evaluation in previous work, I proposed a method based on text readability. Assuming that ‘theme first’ preference is at work in written English (following Mathesius [1975, p. 81], [Halliday, 1967], and [Kuno, 1978] in a slightly weaker form), I adopted the FSP-type approach that a pattern of “*Theme – Rheme*” is more readable than one of “*Rheme – Theme*”. Although certain effects have been observed, the paper did not provide an objective way of measuring the effects. As mentioned in Subsection 2.3.3, ‘theme first’ preference is controversial. The present work does not assume this position in any strong form.

Identification Problem in Natural Language Generation

Natural language (NL) generation is one area where theories of information structure are successfully applied. This involves contextually appropriate generation of intonation in English [Prevost and Steedman, 1993; Prevost, 1995; Prevost, 1996] and that of word order in Turkish [Hoffman, 1994; Hoffman, 1995; Hoffman, 1996].²¹ Such approaches are possible due to direct linguistic marking of information structure. Although the Identification Problem in its original form is not a part of NL generation, there are some connections between them.

First, an assumption common to the above-mentioned NL-generation approaches (except for Hoffman [1996], which also presents an information-structure identification algorithm) is that the information structure is available for each utterance in the given contexts. Prevost [1995] also

²¹Günther et al. [1999] is another example.

works on short discourse, but his examples are limited to the cases where the subsequent utterances share the same theme as the first one. Thus, while usefulness of information structure for NL generation tasks is demonstrated, the question about how information structure generally works in texts is left unanswered.

Now, let us consider the case of generating realistically complex texts. Is the information structure readily available for each utterance? Modern NL generation systems have planning process at the level of content generation as well as surface generation, e.g., McKeown [1985] and Prevost [1995]. Since a typical planner involves propositions as a unit of processing, it may be able to determine the information structure of a complex utterance involving a subordinate clause based on how the utterance is derived in connection to the context. But, since an information-structure division generally corresponds to units smaller than a clause, a process of identifying information structure is still needed.

For the case of a NL generation module as a part of a machine translation system, it is in general impossible that an automated system can derive the ‘intention’ of the writer of the source text, cf. planning in NL generation. In fact, most of the currently available systems simply transfer either isolated syntactic and/or semantic structures between the corresponding utterances. Thus, while a generation module requires a solution to the Identification Problem, the current solutions to NL generation problems involving information structure do not solve the Identification problem.

Note about Evaluation Methodology

As we have seen above, evaluation is a missing component in all previous proposals for the Identification Problem. Let us briefly discuss the methodology we might use for this purpose. One possible direction is to identify a non-linguistic observable phenomena practiced in, e.g., psycholinguistics. They control referential status of physical objects and observe the relation with linguistic expression [Arnold et al., 1997]. But, since we want to evaluate identification processes, this approach does not seem to be applicable to our case. Another technique is to directly observe processing load through eye tracking [e.g., Rayner and Pollatsek, 1987 (a review)]. This seems like a promising possibility, but is beyond the scope of the current work. The present work pursues a purely linguistic way of evaluation in the remainder of this thesis.

2.5 Summary

The main conclusion of this chapter is that the Identification Problem still remains wide open. In the previous section, we identify problems specific to the computational approaches to the Identification Problem. But, more importantly, this group of work lacks the essential properties required for a solution to the problem, i.e., realistic coverage, an evaluation method, and a clear theory-procedure relation.

On the other hand, previous theories of information structure reviewed in Section 2.3 are mostly indifferent to the Identification Problem. Although various properties related to information structure have been investigated, previous theories do not delineate the properties of theme and rheme and the relation between theme and rheme as pursued in our simple hypothesis (20). This situation calls for a theory of information structure that can overcome these problems.