

# Information Structure and Inference Process Involving Inferrable Referents

Nobo Komagata

Department of Computer Science

The College of New Jersey

PO Box 7718, Ewing, NJ 08628

komagata@tcnj.edu

## Abstract

Computationally processing inference is one of the major outstanding problem, and virtually all existing approaches are very expensive. Even a special case of identifying “inferrable” referents is no exception. Among other applications including anaphora resolution, inferrables play an important role in identifying “information structure,” organization of sentence components based on informativeness. Thus, some kind of inference process would greatly affect the intricate problem of identifying information structure and inferrables. Without knowing exactly when to start such an inference process, a computational procedure of identifying information structure would suffer. This paper identifies the condition for starting the inference process to identify inferrables while identifying information structure, which could lead to a reduced processing time for this type of computational analysis.

## 1 Introduction

Inference is one of the most powerful mechanisms involved in language use, which has been tackled from a variety of angles [e.g., Grice, 1975; Hobbs, 1979; Charniak, 1973]. One major problem with inference process is that it is difficult to identify when to start and when to stop the process. For example, Grice discusses a way to start an inference process, referring to violation to his “cooperation maxims.” This and other ideas are still actively debated [e.g., Dale and Reiter, 1996].

At the referent level, some inference process is essential for identifying “inferrable” referents [Prince, 1981]. For example, after being told about a house, one can infer the existence of a door (of that house). Inferrables play an important role in anaphora resolution and the notion of “information structure”, organization of sentence components based on their informativeness [Vallduví, 1990]. If a sentence contain an inferrable, e.g., the door of a house, the inferrable can be understood as a less informative part (called “theme”) and serve as a starting point to provide new information (called “rheme”). This situation can be seen in the following example.

(1) *Q*: How is the house?

*A*: [The door]<sub>Theme</sub> [is purple]<sub>Rheme</sub>.

The response is divided into “theme” and “rheme,” the two components of information structure. Note that these terms are adopted in this paper, following Halliday [1967],<sup>1</sup> mainly to avoid more overloaded pairs such as “old”/“new” and “topic”/“focus”.

Identification of information structure is essential for many Natural Language Processing (NLP) applications. For example, text-to-speech systems could place naturally-sounding pitch accents based on

<sup>1</sup>However, we do not follow Halliday’s position that the theme always comes before the rheme.

information structure [Prevost and Steedman, 1993]. Machine Translation systems could identify special context-dependent particles in the sentences in, say, Japanese translated from English, applying the observation of Kuno [1972] and others. Computer-Assisted Writing systems could evaluate text readability.

However, as seen in example (1), the theme may be an inferrable, which requires some inference process to identify the connection to the context. However, since inference process is in general extremely expensive, we should use it only if it is absolutely necessary. In many cases, we can identify themes without invoking an inference process, e.g., the referent for the theme may have appeared in the discourse explicitly. Then, the main question of this paper is as follows: when exactly should we invoke an inference process to identify an inferrable while identifying information structure?

This paper responds to this question as follows. A reasonably complete computational process can identify all the themes except for inferrables. Thus, in order to identify information structure, we need to invoke an inference process only after we identify non-inferrable themes. In support, we will examine a computational procedure for identifying information structure in real expository texts, the results of the identification process, and its evaluation.

While the usefulness of information structure has been addressed by many researchers, there are many outstanding issues including: its definitions, components, division, etc. Identifying information structure in real texts is a challenge. Thus, this paper will by no means be able to address all the issues raised by the researchers. We will focus on a single point, i.e., the information-structure status of the matrix clause subject.

We will also make an assumption that when a text is translated into another language, the information structure for each sentence would be preserved. Since information structure is realized differently across languages, this assumption will let us use translation to evaluate computational procedures from different perspectives.

For concreteness, this paper adopts a specific view of information structure, a specific way of identifying them, and a specific evaluation method. However, the present discussion and experiment must be repeatable with minor adjustments.

The rest of this paper is organized as follows. Section 2 describes the notion of information structure needed for this paper, as well as a brief description of how information structure is realized across a few languages. Section 3 introduces a computational procedure to identify information structure in real texts. The section also describe the evaluation method, the results, and a detailed analysis of the data.

## 2 Notion of Information Structure

As mentioned in Section 1, there are diverse views about information structure. This section clarifies the position of this paper so that the present approach is applicable to other cases, possibly with modification.

Here, the notion of information structure is characterized by the properties listed below, in principle following the idea in Steedman [2000a]. These properties are at least partially compatible with many definitions.

- (2) *a.* Partition type: Information structure is a binary partition between a theme and a rheme, cf. trinomial [Vallduví, 1990] and graded partitions [Firbas, 1964].
- b.* Relation between the two components
  - (i) The composition of the semantic representations of the theme and the rheme yields the proposition, i.e.,  $(theme)(rheme) = proposition$ .
  - (ii) A rheme is explicitly required, but a theme is only implicitly required (i.e., may not be realized linguistically).
- c.* Linguistic constraints on the two components (theme and rheme)

- (i) The components may be marked in a variety of linguistic forms.
  - (ii) The components may span more or less than traditional constituents [Steedman, 2000a].
  - (iii) The components may correspond to various abstract categories (related to the ideas in Asher [1993]).
  - (iv) The partition of the two components is not always clear cut.
- d. Contextual constraints on the components: The theme must be “linked” to the context. The rheme can, but not necessarily be linked.

These properties will be referred to in the description of the procedure to identify information structure.

### 3 Identifying Information Structure

In this section, we review a sample process of identifying information structure in real texts and show that we can identify virtually all the themes except for inferrables. Although there have been several proposals to identify information structure [e.g., Hahn, 1995; Hajičová et al., 1995; Hoffman, 1996], this paper focuses on Komagata [1999] because it is the only one that deals with real texts and also discusses evaluation results.

The experiment uses a set of texts consisting of medical case reports from a journal, *The Physician and Sportsmedicine*. In order to evaluate generalizability of the procedure, the original experiment divided texts into two groups: the training data (16 texts) and the reserved test data (8 texts). This aspect is justified in the original experiment, and will not be repeated here. Thus, this paper will deal with all 24 texts (197 sentences) together.

The identification process adopts an analysis integrating both contextual and linguistic aspects. To be able to deal with realistic partitions of information structures, including “non-traditional constituents” in property (2-cii), the procedure depends on a parser based on Combinatory Categorical Grammar [Steedman, 2000b].

#### 3.1 Criteria for Contextual Linking: Theme Candidacy

The most crucial aspect of the process is to identify theme candidates. Based on property (2-d), a theme must be linked to the context (called “contextual link”), either linguistically and extra-linguistically. Further dividing the extra-linguistic condition, we consider the following criteria for identifying contextual linking.

- (3) a. Linguistic
  - b. Extra-linguistic
    - (i) Discourse-oldness
    - (ii) Domain-specific knowledge

Note that linguistic and discourse-oldness are analyzed recursively along the bottom-up parsing process. Although these criteria were partially confirmed in the experiment, they are not claimed to be complete; further refinements would improve the accuracy. In the following, the three criteria will be described more in detail.

First, linguistic marking is analyzed recursively, dealing with various linguistic units in a bottom-up manner. The following linguistic marking either sets or resets the contextual linking status of the current constituent.

- (4) a. Definite expressions: Definite article (*the*), definite pronoun (e.g., *this*), and demonstratives (e.g., *this*) generally set the contextual linking status. Exceptions include the “logical” use of a definite expression (e.g., *the first bus*) [Quirk et al., 1985], which can be detected linguistically. Note that there are exceptional cases.

- b. Sentence-initial modifiers: Sentence-initial adverbial phrases (e.g., *until the early 1980*) and subordinate clauses (e.g., *as it is used here*) set the contextual linking status.
- c. Indefinite expressions: Indefinite article (*a/an*) and indefinite pronouns (e.g., *anyone*) generally reset the contextual linking status. Exceptions include two-place nouns (e.g., *brother*) and indefinite expression with a discourse-old head (considered as a generic reference), both of which suggest some connection to the context.

The following group of linguistic marking projects the contextual linking status of the head or argument. The information is mainly derived by examining the training data of this experiment, and in general confirmed in the test data as well.

- (5) a. Non-definite determiners (e.g., *many*), argument-taking pronouns (e.g., *many of*), and auxiliary verbs (e.g., *will*) project the status of their argument.
- b. Restrictive post-nominal modifiers (e.g., *of tuberculosis*) project its status to the unit spanning from the noun to the post-nominal modifier phrase.
- c. Coordination (e.g., *proprioceptive training and proprioceptive rehabilitation*) projects the contextual linking status if both conjuncts are contextually linked.
- d. Noun-noun compounds (e.g., *exercise program*) project the status of the first noun.
- e. Denominal adjectives (e.g., *medical*), closely related to a noun, project its status.
- f. Regular adjectives (e.g., *active*) project the status of the noun.
- g. Possessives (e.g., *a patient's*) project the status of the possessor.

There are other linguistic marking that can be used for identifying theme candidates, e.g., topicalization/focus movements, pseudoclefts, etc., but they were not found in the data and not implemented in the experiment.

Second, discourse-oldness is captured as a repetitive occurrence of semantic representations stored in a generalized “context set” (originally proposed as a set of propositions [Stalnaker, 1978]). The proposed context set contains semantic representations corresponding to all linguistic units involved in the successful parses. The process of identifying discourse-oldness status is thus a search for a semantic representation through this context set. Since the context set is updated every time a sentence is processed, discourse-oldness is limited to the inter-sentence case; no intra-sentence linking can be detected. Note that the search process does not involve inflectional information, allowing a match between, e.g., “physician” and “physicians.”

Third, domain-specific knowledge is handled as a variation of discourse-oldness. That is, certain semantic representations are added to the context set at the very beginning of the process, in order to simulate the availability of extra-linguistic elements. In the experiment dealing with medical case reports, the lexical entry “physician,” “clinician,” and “patient” are inserted in the context set at the time of initialization. Then, occurrences of these words (and their inflectional variations) are analyzed as discourse-old, and thus as contextual links.

### 3.2 Identifying Theme-Rheme Partitions

The process of identifying information structure proceeds in a bottom-up manner along with parsing. Whenever a linguistic unit is formed, (1) the relevant linguistic marking is identified, if any, and (2) its semantic representation is checked for discourse-oldness against the context set. These criteria for contextual linking are specified at the local level and applied recursively to larger linguistic units. Thus, the process can apply to arbitrary linguistic structure consisting of the defined basic components. At the final step of parsing, the last semantic composition is examined for information structure. If the two units are a pair of a contextual

link and a non-contextual link, these units are identified as the theme and the rheme, respectively. If both of the units are non-contextual links, the entire utterance is identified as a rheme. If both of the units are contextual links (i.e., both theme candidates), one must be the rheme because there must be a rheme as required by property (2-bii). In this case, the current procedure chooses the theme-rheme ordering by default, reflecting the general tendency of that ordering.

Here is an example of identifying information structure with some description for each sentence (part of Text 12).

(6) *i.* (Title) Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

*Note: Titles are not analyzed for their information structure because they do not correspond to full proposition. Other sentences excluded from information-structure analysis include a parenthetical for a complete sentence and quotes.*

- ii.* [Osteoporosis]<sub>Theme</sub> [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”]<sub>Rheme</sub>.”

*Explanation: The subject, osteoporosis, is identified as a contextual link due to its discourse-oldness. None of the remaining part give rise to the contextual link status. When the semantic composition of the subject and the VP is formed, these components are labeled as theme and rheme, respectively.*

- iii.* [Although anyone can develop osteoporosis]<sub>Theme</sub>, [postmenopausal women and young females with menstrual irregularities are most commonly affected]<sub>Rheme</sub>.

*Explanation: The sentence-initial subordinate clause is identified as a contextual link. Since there are no other contextual links available, when the subordinate and the main clauses are composed, the above information structure is identified.*

- iv.* [An estimated 20% of women more than 50 years old have]<sub>Rheme</sub> [osteoporosis]<sub>Theme</sub>.

*Explanation: The only contextual link is the object, osteoporosis. Since the current analysis accepts the above, non-traditional division, information structure is analyzed accordingly.*

- v.* (continued)

### 3.3 Evaluation

The evaluation process is based on the idea that information structure is maintained through translation. When a text in English is translated into Japanese, the theme-rheme status of the matrix subject is in general identifiable by particle choice in Japanese. In general, if the matrix subject is a part of the theme, the subject noun phrase will be marked with *wa* (thematic marker). Otherwise, it will be marked with *ga* (nominative case marker). Although this distinction is sufficient for the present evaluation purposes, the situation is not that simple. More details are discussed in my dissertation [Komagata, 1999].

In the following, the matrix subjects that fall within the theme are predicted to be marked with *wa* and those within the rheme, with *ga*, in the corresponding translation in Japanese.

(7) *i.* (Title) Osteoporosis in Active Women: Prevention, Diagnosis, and Treatment

- ii.* [Osteoporosis]<sub>Theme</sub> *wa* [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”]<sub>Rheme</sub>.”

- iii.* [Although anyone can develop osteoporosis]<sub>Theme</sub>, [postmenopausal women and young females with menstrual irregularities]<sub>Rheme</sub> *ga* are most commonly affected]<sub>Rheme</sub>.

Sentence number	Machine prediction	Translator			Target choice	Evaluation status
		<i>wa</i>	<i>ga</i>	other		
(ii)	<i>wa</i>	4	0	0	<i>wa</i>	Correct
(iii)	<i>ga</i>	0	1	3	n/a	Not confirmed
(iv)	<i>ga</i>	0	3	1	<i>ga</i>	Correct
(v)	<i>wa</i>	4	0	0	<i>wa</i>	Correct
(vi)	<i>wa</i>	2	0	2	<i>wa</i>	Correct
(vii)	<i>wa</i>	4	0	0	<i>wa</i>	Correct
(viii)	<i>ga</i>	0	4	0	<i>ga</i>	Correct
(ix)	<i>wa</i>	4	0	0	<i>wa</i>	Correct

Table 1: Comparison of Machine Prediction and Translator’s Particle Choices

Category	Predicted	Correct	Error	Not confirmed
Theme	128	104 (81%)	2 (2%)	22 (17%)
Rheme	42	11 (26%)	16 (38%)	15 (36%)
Overall	170	116 (68%)	18 (10%)	37 (22%)

Table 2: Evaluation Results

iv. [An estimated 20% of women more than 50 years old ga have]<sub>Rheme</sub> [osteoporosis]<sub>Theme</sub>.

v. (continued)

The experiment recruited 4 translators for the training data and 6 translators for the test data. For the purpose of evaluation, we also prepare a set of target particle choices representing a hypothetical translator in the following way. If either *wa* or *ga* is chosen by at least two more translators than the other, that particle is considered as target.<sup>2</sup> If the difference is one or zero, neither *wa* nor *ga* is chosen as the target. In Japanese, subjects can be dropped if they can be recovered from the context. In this case, the subject is considered as a part of the theme and classified along with *wa*-marking. Note that it is also possible that the matrix subject is marked with particle other than *wa* or *ga*, reflecting structure change. This is another case where neither particle is chosen as the target. The agreement among translators was acceptable, but not extremely good. A detailed analysis using the  $\kappa$  statistic is included in Komagata [1999].

### 3.4 Results

First, the details of Text 12 evaluation are shown in Table 1. The results for the entire 24 texts are summarized in Table 2. A fair number of instances are labeled as “not confirmed.” This category involves two cases. In the first case (27 out of 37 non-confirmed instances), the translators’ use of the particles were not consistent, i.e., not sufficient discrimination, corresponding to property (2-civ). In the second case (10 out of 37), neither of the two particles were chosen due to the structural change during translation. Since neither of these cases confirm the machine prediction positively or negatively, we conclude that the theme identification is fairly accurate with only 2% error.

The main problem lies with the accuracy of identifying rheme, with 38% error. Earlier, we noted that this is mainly due to the presence of “inferrables.” To verify this point, we will next examine all the error cases.

<sup>2</sup>This method of target selection is different from Komagata [1999].

The complete data in various forms are contained in the following files available on-line at: “<http://www.tcnj.edu/~komagata/pub/IS-Data>”.

- Original texts: Text1Original.pdf
- Mechanical prediction of particle choices: Text2Predicted.pdf
- Evaluation compared with human translators: Text3Evaluated.pdf
- Annotated information structure analysis: Text4Annotated.pdf and Text4Annotated.txt
- Data summary file: IS-Data.xls

### 3.5 Data Analysis and Discussion

In this subsection, we analyze all the errors and conclude that the proposed system is capable of identifying practically all the theme cases. If that is the case, we can support the paper’s claim that the failure to identify theme cases would trigger the inference process to identify inferrables. In the following, we will focus on the following types of errors: (i) predicted theme, but actually rheme and (ii) predicted rheme, but actually theme. Note that the  $j$ th sentence in Text  $i$  will be shown below as  $(i-j)$ .

In the data, there are two instances where the rheme was incorrectly predicted as theme. The first one (4-5) is shown below.

- (8) In addition, the incidence of multiple-drug-resistant *Mycobacterium tuberculosis* has increased and made treatment more difficult.

The analysis in Komagata [1999] was as follows. The conjoined predicate involves a “stage-level” predicate (roughly, a predicate to describe a temporary state, cf. “individual-level” [Carlson, 1980]). A stage-level predicate must have an event argument [Kratzer, 1995] and it would be difficult to form a theme spanning both the event argument and the subject. Then, the subject must belong to the rheme and must be marked with *ga*. If this is the case, it would be possible to identify the information structure involving a stage-level predicate because it is in general possible to detect a stage-level predicate. Note that in many cases, including the above example, the event argument is not linguistically realized.

However, it is problematic to state that the (matrix) subject of a stage-level predicate always belongs to the rheme. For example, the following sentence (4-4), immediately preceding (8) above has a subject, which is a part of the theme (marked with *wa* in the translation).

- (9) However, the incidence of the disease in the United States, including those who had tuberculosis infection as well as the active disease, increased by approximately 14% from 1985 to 1993, largely because of the effects of homelessness, alcoholism, drug dependency, immigration from endemic areas, and the human immunodeficiency virus (HIV) epidemic.

In this example, the event is linguistically realized after the verb and is a part of the rheme. A possible revision of the hypothesis in Komagata [1999] is as follows: the subject of a stage-level predicate is a part of the rheme if the event argument is not linguistically expressed (i.e., a part of the theme).

The second problematic instance (18-6) involves a complex indefinite subject.

- (10) an understanding of SUI and the wide range of available treatments is important for fitness-oriented physicians

The procedure first detects the discourse-old status of *SUI* and the definiteness of *the wide range of available treatments*. The coordination of these conjuncts thus results in a contextual link. This contextual-link status is projected through the preposition *of*, to the noun and the formation of the prepositional phrase. The composition of an indefinite article with the contextual link is analyzed as a generic, and thus set a contextual-link status, based on criterion (4-c). This puts the subject as a part of the theme, and predicts

the use of *wa*. This suggests that criteria (4-c), i.e., our conjecture about the indefinite generic, needs to be re-examined.

There are three different types of incorrect rheme predictions. The first type is an idiosyncratic situation involving coordination (16-4), as shown below.

(11) Among athletes, ankle sprains are the most common injury, and inversion injuries are frequent.

The system predicted *ga* based on the fact that the two clauses are coordinated and that the latter clause is all rheme. Thus, the result of coordination ended up with all rheme. If this sentence did not have the latter clause, the subject *ankle sprains* would be analyzed as discourse-old, referring to an earlier instance of *ankle sprain* and could have predicted the use of *wa*.

The second type of errors involve discourse-initial accommodation (6-2, 9-2, 16-2, 21-2, 23-4, 24-2). In all 24 texts, the translation of the first sentence after the title employs *wa*-marking on the matrix subjects. This suggests that discourse-initial accommodation is quite strong. Thus, although this point is beyond the sentence-level linguistic analysis adopted in the system, it is possible to mechanically process these cases at the discourse level.

The third type of errors involve inferrable cases, which is by far the most common. The following (3-5) is a typical situation.

(12) *i.* Cheerleading competitions are held at regional and national levels,  
*ii.* and training is a year-round activity.

Here, the underlined word *training* is inferrable from the phrase *cheerleading competitions*, but *training* is neither discourse-old nor linguistically-marked as a candidate for the theme. Without an inference mechanism, this type of error is unavoidable.

The other eight cases are shown below.

- 2-3: “A fiberglass cast with a waterproof liner that “breathes” ” inferrable from “waterproof cast”
- 5-4: “Musculoskeletal weakness, stiffness, and pain” inferrable from “decreased mobility”
- 14-3: “Exercise-related symptoms in the upper GI tract” inferrable from “athletes”
- 19-10: “aging tissues” inferrable from “adult recreational athletes”
- 20-3: “Youth athletic programs” inferrable from “children and adolescence” and “exercise”
- 20-4: “Peer socialization” inferrable from “youth athletic program”
- 20-7: “Diagnostic and treatment efforts” inferrable from “injuries”
- 21-3: “Employers” inferrable from “workplace”

Going through all the error cases, we can conclude that the main problem that goes beyond the current technique is with inferrables, involving an inference process.

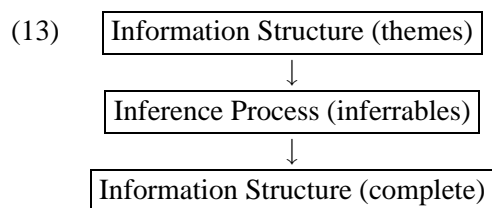
## 4 Conclusion

Referring to a recent experiment, this paper confirms that information structure is computationally identifiable in real texts based on linguistic and limited contextual analyses, *except for inferrables*. In order to completely identify information structure, we will need to have some inference process. However, we will need to invoke the process only for the cases where inferrables may be involved. In the experiment data, only the 42 instances (out of 170) which are labeled as rheme may contain inferrables. Thus, the inference process will need to analyze these cases only. That is, we will need to invoke an inference process (for

identifying information structure) for approximately 25% of all the sentences. The situation has some similarity to but is still different from the use of Gricean Maxims of Cooperation as a starting point for general inference because in order to detect “flouting” instances, we will already need some inference process.

Let us compare the above situation with the following. A typical approach to English-Japanese machine translation would label all the matrix subjects with *wa*, with no automatic mechanism of identifying information structure. Such a system can outperform the experiment examined here by a small margin because of the theme-first tendency. However, suppose that an inference system to identify inferrables is available, still with high operating cost. Not knowing which matrix subjects surely belong to a theme, such a system will need to invoke an inference process 100% of the time to choose appropriate particles.

While this paper focuses on the starting point of an inference process, we could also extend our discussion to the ending point. Now, the process of identifying information structure cannot be complete without an inference process. That is, the end point of this type of inference process must depend on whether or not an inferrable can be identified as a part of the theme. However, it is not clear when such a process should terminate. This may be the reason why there are not many of these cases. For example, it would be less costly to use definiteness to signal an inferrable. Reflecting the above discussion, we may hypothesize that information-structure analysis and inference process cannot be ordered sequentially (shown below).



There is another suggestion that information structure and discourse structure must be analyzed in parallel [Komagata, 2003]. Then, a NLP system must integrate and operate these modules in parallel.

One way to extend the present approach to identifying information structure is to use linguistic marking in more diverse languages. For example, translation into Czech would provide the word order as linguistic marking of information structure [Firbas, 1964].

One of the problems with pragmatic analysis and processing is that situations can be black-and-white property (2-civ). While it is important to analyze clear cases, it might also be helpful to analyze murky cases. For example, what would happen to the speaker and the listener if they cannot resolve an inferrable? Would it be possible to characterize clear cases as special cases of all the cases? Responses to these questions might lead to an approach that transcend the symbolic vs. non-symbolic distinction.

## Bibliography

- Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer.
- Greg N. Carlson. 1980. *Reference to kinds in English* (originally a PhD thesis in 1977). Garland.
- Eugene Charniak. 1973. Jack and Janet in Search of a Theory of Knowledge. In *Advanced Papers from the Third International Joint Conference on Artificial Intelligence, Stanford, CA*, pages 337–343. Morgan Kaufmann.
- Robert Dale and Ehud Reiter. 1996. The Role of the Gricean Maxims in the Generation of Referring Expressions. In *AAAI Spring Symposium on Computational Models of Conversational Implicature*.
- Jan Firbas. 1964. On Defining the Theme in Functional Sentence Analysis. *Travaux Linguistiques de Prague*, 1:267–280.

- H. P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics, 3: Speech Acts*, pages 305–315. Academic Press.
- Udo Hahn. 1995. Distributed Text Structure Parsing – Computing Thematic Progression in Expository Texts. In Gert Rickheit and Christopher Habel, editors, *Focus and Coherence in Discourse Processing*, pages 214–250. Walter de Gruyter.
- Eva Hajičová, Hana Skoumalová, and Petr Sgall. 1995. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics*, 21(1):81–94.
- Michael A. K. Halliday. 1967. Notes on Transitivity and Theme in English (Part II). *Journal of Linguistics*, 3:199–244.
- Jerry R. Hobbs. 1979. Coherence and Coreference. *Cognitive Science*, 3:67–90.
- Beryl Hoffman. 1996. Translating into Free Word Order Languages. In *COLING-96*, pages 556–561.
- Nobo N. Komagata. 1999. *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese*. PhD thesis, University of Pennsylvania.
- Nobo Komagata. 2003. Information Structure in Subordinate and Subordinate-like Clauses. *Journal of Logic, Language and Information*, 12(3):301–318.
- Angelica Kratzer. 1995. Stage-level and Individual-level Predicates. In Gregory N. Carlson and Francis Jeffrey Pelletier, editors, *The Generic Book*, pages 125–175. University of Chicago Press.
- Susumu Kuno. 1972. Functional Sentence Perspective: A Case Study from Japanese and English. *Linguistic Inquiry*, 3(3):269–320.
- Scott Prevost and Mark Steedman. 1993. Generating Contextually Appropriate Intonation. In *EACL 6*, pages 332–340.
- Ellen F. Prince. 1981. Toward a Taxonomy of Given-New Information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.
- Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics, Vol. 9: Pragmatics*, pages 315–322. Academic Press.
- Mark Steedman. 2000a. Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry*, 31(4):649–689.
- Mark Steedman. 2000b. *The Syntactic Process*. MIT Press.
- Enric Vallduví. 1990. *The informational component*. PhD thesis, University of Pennsylvania.