

A COMPUTATIONAL ANALYSIS OF INFORMATION STRUCTURE
USING PARALLEL EXPOSITORY TEXTS
IN ENGLISH AND JAPANESE

Nobo N. Komagata

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1999

Dr. Mark J. Steedman
Supervisor of Dissertation

Dr. Jean Gallier
Graduate Group Chair

COPYRIGHT

Nobo N. Komagata

1999

Acknowledgments

I would like to thank my advisor, Mark Steedman, for his valuable advice, warm encouragement, and unfailing patience throughout my years at Penn. His extremely broad interests and deep insight have always guided my work on this thesis, especially at difficult times. Among many things I learned from Mark, I particularly appreciate his points that I needed to deliver results visibly and demonstrate the generality of an idea.

I am deeply grateful to my thesis committee members. Claire Gardent made critical comments that were difficult to respond to but gave me an opportunity to look at the thesis from a different point of view. Aravind Joshi inspired me on a wide range of issues, from formal grammars to discourse analysis. Martha Palmer read a draft of this thesis with great care and made numerous helpful points about the contents. Bonnie Webber gave me an opportunity to assist in her enjoyable AI course for three semesters. Although Ellen Prince is not on the thesis committee, I thank her for serving on the proposal committee and teaching me linguistic pragmatics.

The superb academic environment at Penn is one of the factors I cannot forget. I learned a great many basic concepts from the coursework in computer science and linguistics. I thank my professors, especially Peter Buneman, Robin Clark, Tony Kroch, Mitch Marcus, and Max Mintz. I would like to thank Mike Felker for helping me at every stage, from application to graduation. In addition, I owe much to the people who made possible the financial support I received in the form of a Dean's Fellowship, CIS departmental grants, and an IRCS Graduate Fellowship.

I am grateful for my colleagues at Penn. In particular, I had valuable comments from the members of Mark Steedman's research group, especially Beryl Hoffman, Charlie Ortiz, Jong Park, Scott Prevost, Matthew Stone, and Mike White. At various stages, I also received helpful comments from Jason Baldridge, Susan Converse, Miriam Eckert, Jason Eisner, Chung-hye Han, Gerhard Jäger,

Seth Kulick, Rashmi Prasad, Anoop Sarkar, Bangalore Srinivas, and Michael Strube. Thanks also to Mimi Lipson for her help in designing a linguistic experiment. In addition, discussion with Penn graduates and visitors was a great help; many thanks to Sadao Kurohashi, Kathy McKeown, K. Vijay-Shanker, and David Weir.

My interest in computational linguistics began with my undergraduate project. I thank Tetsuya Okamoto for introducing me to the field and supervising the project. Later, I had a wonderful opportunity to be involved in a machine-translation project at Bravice with Naoya Arakawa, Akira Nagasawa, and Jun Ohta. Special thanks to Akira Kurahone, who introduced me to Categorical Grammar, the grammatical framework used in this thesis.

I would like to thank a very special couple, my mentors, Josephine and José Rabinowitz, for their encouragement, support, and the best Mexican food in town right from the next door.

Finally, I am very fortunate to have had the understanding and patience of my family from distant Tokyo, Sydney, and Québec. I thank my parents, Ichiko and Eiichi, for the way they brought me up and what they have given to me. But most importantly, my hearty thanks to my wife, Sachiko, for everything we have shared since 1983. While her expertise with medical terminology and statistics, especially the kappa statistic, was an invaluable help, the single most important thing for me has been and always will be her smile.

Abstract

A COMPUTATIONAL ANALYSIS OF INFORMATION STRUCTURE USING PARALLEL EXPOSITORY TEXTS

IN ENGLISH AND JAPANESE

Nobo N. Komagata

Supervisor: Dr. Mark J. Steedman

This thesis concerns the notion of ‘information structure’: informally, organization of information in an utterance with respect to the context. Information structure has been recognized as a critical element in a number of computer applications: e.g., selection of contextually appropriate forms in machine translation and speech generation, and analysis of text readability in computer-assisted writing systems.

One of the problems involved in these applications is how to identify information structure in extended texts. This problem is often ignored, assumed to be trivial, or reduced to a sub-problem that does not correspond to the complexity of realistic texts. A handful of computational proposals face the problem directly, but they are generally limited in coverage and all suffer from lack of evaluation. To fully demonstrate the usefulness of information structure, it is essential to apply a theory of information structure to the identification problem and to provide an evaluation method.

This thesis adopts a classic theory of information structure as binomial partition between theme and rheme, and captures the property of theme as a requirement of the contextual-link status. The notion of ‘contextual link’ is further specified in terms of discourse status, domain-specific knowledge, and linguistic marking. The relation between theme and rheme is identified as the semantic composition of the two, and linked to surface syntactic structure using Combinatory

Categorial Grammar. The identification process can then be specified as analysis of contextual-link status along the linguistic structure.

The implemented system identifies information structure in real texts in English. Building on the analysis of Japanese presented in the thesis, the system automatically predicts contextually-appropriate use of certain particles in the corresponding texts in Japanese. The machine prediction is then compared with human translations. The evaluation results demonstrate that the prediction of the theory is an improvement over alternative hypotheses. We then conclude that information structure can in fact be used to improve the quality of computational applications in practical settings.

Contents

Acknowledgments	iii
Abstract	v
Notational Conventions	xiv
1 Introduction	1
2 Information Structure: The State of the Art and Open Questions	14
2.1 The Identification Problem	14
2.2 What is Information Structure?	16
2.3 Previous Theories of Information Structure	23
2.3.1 Referential Status of Theme and Rheme	24
2.3.2 Information Structure vs. Contrast	31
2.3.3 Information Structure and Linguistic Form	36
2.3.4 Internal Organization of Information Structure	40
2.4 Previous Proposals for Identifying Information Structure	45
2.5 Summary	53
3 A Theory of Information Structure	54
3.1 Main Hypothesis: Semantic Partition between Theme and Rheme	54
3.2 Contextual Link	58
3.2.1 Contextual Link and Inference	58
3.2.2 Logic-External Properties for Bounding Inference	60

3.3	Linguistic Marking in English	62
3.3.1	Linguistic Marking for Contextual Links	63
3.3.2	Special Constructions	74
3.4	Grammatical Components	80
3.4.1	Syntax-Semantics Interface	81
3.4.2	Flexible Constituency	83
3.5	Discontiguous Information Structure	85
3.6	Summary	90
4	Formalization of the Theory with Combinatory Categorical Grammar	92
4.1	Combinatory Categorical Grammar	92
4.1.1	Motivation	93
4.1.2	Derivation Examples	95
4.1.3	Standard CCG: A Summary	98
4.1.4	Extensions of CCG	100
4.1.5	Generative Power and Theoretical Parsing Efficiency	103
4.2	Specification of Contextual-Link Status	105
4.3	Integration of Structured Meaning	109
4.3.1	Composition of Structured Meanings	109
4.3.2	Identification of Information Structure	116
4.3.3	Analysis of Gapping	116
4.4	Summary	120
5	Realization of Information Structure in Japanese	121
5.1	Introduction	121
5.2	Functions of Particle <i>wa</i>	125
5.2.1	Two Functions of <i>wa</i>	126
5.2.2	Contrastive Function	127
5.2.3	Thematic Function	133
5.3	Function of Long-Distance Fronting	140
5.4	Prediction of <i>wa</i> and <i>ga</i> from Information Structure	143

5.5	Summary	149
6	Implementation of the Information-Structure Analyzer	150
6.1	Introduction	150
6.2	Practical CCG Parser	152
6.2.1	Requirements for the Parser	152
6.2.2	Elimination of Spurious Ambiguity	153
6.2.3	Linguistic Specification and Processing	155
6.2.4	Performance	163
6.3	Processing Information Structure	165
6.3.1	Discourse Status and Domain-Specific Knowledge	165
6.3.2	Linguistic Marking of Contextual Links	167
6.3.3	Composition of Structured Meaning	170
6.3.4	Identification of Information Structure	174
6.3.5	Prediction of Particle Choice in Japanese	175
6.3.6	Potential Applications to Generation	178
6.4	Summary	179
7	Evaluation of the Theory Using Parallel Texts	181
7.1	The Data	181
7.2	Development of an Evaluation Method Using the Training Data	183
7.2.1	Mechanical Prediction of Particle Choices in Japanese	184
7.2.2	Human Translation	186
7.2.3	Evaluation Methodology	190
7.2.4	Analysis of Errors	193
7.2.5	Possibility of Extending the Evaluation	195
7.3	Evaluation of the Theory Using the Test Data	198
7.3.1	Extension of the System for the Test Data	198
7.3.2	Results	199
7.3.3	Discussion	201
7.4	Summary	204

8	Conclusion	205
A	Generative Power and Parsing Efficiency of CCG-GTRC	211
A.1	CCG with Generalized Type-Raised Categories	211
A.2	Weak Equivalence of CCG-GTRC and CCG-Std	221
A.3	Worst-Case Polynomial Recognition Algorithm	233
A.4	Progress Towards a Practical Parser for CCG-GTRC	240
A.5	Conclusion	244
	Bibliography	247
	Index	271

List of Tables

1.1	Particle Choices by Translators	3
1.2	Particle Choices and Simple Hypotheses	4
2.1	Taxonomy of Assumed Familiarity (adapted from Prince [1981, 1992])	28
3.1	Corpus Analysis of Clefting [Collins 1991]	77
5.1	Realization of Information Structure in Japanese (preliminary)	125
5.2	Contrastive Function of <i>wa</i>	131
5.3	Subject Marking in Embedded Environments	134
5.4	Contrastiveness and Information Structure for <i>wa</i> at the Matrix Level	137
5.5	<i>wa</i> vs. <i>ga</i> at Embedded Environments	138
5.6	<i>wa</i> vs. <i>ga</i> at the Matrix Level	139
5.7	<i>wa</i> and Case Particles in Embedded Environments	144
5.8	<i>wa</i> vs. <i>ga</i> at the Matrix Level	145
5.9	<i>wa</i> and Case Particles at the Matrix Level	146
7.1	Training and Test Data Set	182
7.2	Particle Choices by Human Translators (Text 12)	188
7.3	Particle Choices by Translators (Training Data)	189
7.4	Agreement between Two Translators (Training Data)	190
7.5	Comparison of Hypotheses on the Training Data	191
7.6	Particle Choices by Translators (Test Data)	200
7.7	Agreement between Two Translators (Test Data)	200

7.8	Comparison of Hypotheses on the Test Data	200
A.1	Combinatory Cases for CCG-GTRC	217

List of Figures

1.1	The Phenomenon under Investigation	7
1.2	Limitations of Previous Approaches to the Identification Problem	8
1.3	Overview of the Project	11
2.1	Text Link	19
2.2	Information Structure vs. Contrast	32
3.1	Syntax and Semantics along Linguistic Structure	83
5.1	Particle Prediction in Japanese	149
6.1	System Architecture	151
6.2	CKY-Style Parsing Table	163
6.3	A Summary of the Procedure to Identify Contextual-Link Status	180
A.1	GTRC Recovery Algorithm	238
A.2	Basic Data Set (linear scale)	242
A.3	Basic Data Set (log scale)	242
A.4	Extended Data Set (linear scale)	242
A.5	Extended Data Set (log scale)	242

Notational Conventions

- *Italic*: (1) Cited word, (2) Grammatical subject (in examples) [p. 2], (3) Utterance number in Roman numerals (in discourse examples) [p. 1]
- *Math font* (appears very similar to *italic*): (1) Semantic representation [p. 81], (2) CCG category [p. 95]
- **Boldface**: (1) Technical term with definition/description, (2) Phonological prominence (in examples) [p. 32]
- Sans serif: Category variable (for type raising) [p. 103]
- SMALL CAPS: (1) Terms for referential status from Prince [1981] [p. 28], (2) Grammatical labels (Japanese)
- Typewriter font: Computer source code, output, or data
- Underline: (1) Attention to an element in examples (not a phonological/pragmatic feature), (2) Cancelled categories in CCG derivations [p. 96]
- Single quotes ‘ ’: (1) Technical terms in general, (2) Special character and short symbol
- Double quotes “ ”: Cited expression (more than one word)
- Parentheses (): (1) Argument of functional application [p. 82], (2) Presupposition (in gloss) [p. 126], (3) Utterance number in the form of ($T - U$) corresponding to the U 'th utterance in Text T [p. 182]
- Square brackets []: (1) Citation, (2) Span of information-structure units (i.e., theme/rheme), (3) Functor of functional application [p. 82]

- Curly brackets { }: Span of coordination
- Angle brackets $\langle \rangle$: (1) Exclusion from information-structure analysis [p. 2], (2), Structured meaning (as in $\langle X, Y \rangle_{L-R}$ where L/R are left/right boundaries) [p. 88], (3) General meta-variable [p. 95]
- Double square brackets $\llbracket \rrbracket$: Semantic value [p. 82]
- Asterisk *: Ungrammatical sentence [p. 17]
- Number sign #: Contextually inappropriate utterance [p. 17]
- Prime '': Translation of linguistic expression to semantic representation [p. 82]
- Right-arrow \longrightarrow : CCG rule [p. 96]
- Hollow circle \circ : Functional composition [p. 97]
- Plus +: Category combination [p. 110]
- Up-arrow (superscript) \uparrow : Type-raised category [p. 97]
- Double slash //: Modification structure (in semantic representations) [p. 81]
- Grammatical labels for Japanese:
 - TOP = topic marker (thematic function of wa)
 - CONT = contrastiveness marker (strong contrastive function of wa)
 - NOM = nominative case marker
 - ACC = accusative case marker
 - DAT = dative case marker
 - GEN = genitive case maker
 - COP = copula
 - COMP = complementizer
 - NML = nominalizer
 - NEG = negation
 - Q = question