

## Chapter 8

# Conclusion

### Summary

In computational applications such as machine translation, speech generation, and writing assistance, the effect of information structure is critical for contextually appropriate processing of natural language. This thesis focuses on the problem of identifying information structure in expository texts.

But, as we review in Chapter 2, the existing analyses of information structure cannot directly be applied to the Identification Problem. They basically do not address the problem, and are not sufficiently explicit for the purposes of formalization and implementation either. The computational proposals directly responding to the problem are mostly not applied to realistic texts and do not provide an evaluation method.

Our response to this situation is to propose an explicit theory of information structure, formalize and implement it, and evaluate the result with respect to an independent observation. In Chapter 3, we develop a theory of information structure with the Identification Problem in mind. The main hypothesis is that information structure is a semantic composition between a theme and a rheme and the theme is necessarily contextually-linked. Following the Montagovian tradition, we analyze instances of semantic composition along the syntactic derivation. This way, the analysis of contextual links in an utterance can be used to identify a information structure of the utterance. The present approach captures the properties of contextual linking in terms of logic-external properties: discourse status, primitive domain-specific knowledge, and linguistic marking. Each of

these properties is precisely described.

For two potential problems with binomial partition of information structure, i.e., non-traditional constituency and discontinuous information structure, we adopt a flexible notion of constituency recognized by Combinatory Categorical Grammar (CCG) and an additional degree of freedom gained by structured meanings compositionally built for CCG constituents as semantic representations.

To establish the connection between the proposed theory and a practical implementation, we formally describe the theory, including the specification of contextual links and structured meanings, within an extended form of the CCG framework (Chapter 4). We also show that variants of CCGs are comparable to the related formalisms with respect to generative capacity and theoretical parsing efficiency.

For the evaluation purpose, we take advantage of the particle choice problem in English-Japanese translation. Chapter 5 provides the basis for this approach by investigating the Japanese particle *wa* and other case markers, and the function of long-distance fronting in detail. After identifying several exceptional cases, we analyze that *wa* and *ga* at the matrix level mark (a part of) theme and rheme, respectively.

The next step is to provide a procedure to identify information structure. In Chapter 6, we first show the practicality of our CCG parser, and then implement the specification of contextual linking and information structure. There are certain procedural aspects associated with our information-structure analysis. These are introduced in a modular fashion, and can be considered reasonable through the examination of the experiment (training) data. As the last step of the mechanical procedure involved in the current project, we apply the analysis of Japanese and predict particle choices for matrix subjects based on the identified information structure.

Finally, the crucial element of this thesis is the evaluation of the theory (Chapter 7). The methodology is to compare the particle predictions made by the system and human translations. We first develop our evaluation method using the training data, and then show that the theory generalizes to previously-withheld data. This demonstrates that the proposed theory is an improvement over the alternative hypotheses underlying the existing computational approaches, and that the proposed theory generalizes to new data.

## Contributions

The main contribution of this thesis is a demonstration, including an evaluation on test materials withheld from the development set, that information structure can be correctly interpreted and used in practical applications such as machine translation for limited domains. This development advances the state where the notion of information structure has rarely escaped the intuition of some researchers.

The first crucial step in this demonstration is to squarely face the Identification Problem. Like other computational approaches to the Identification Problem, but, unlike most theoretical work in linguistics, the current proposal can directly connect the result of the project to practical applications.

The present work is distinguished from other computational approaches in that the results are evaluated based on an independently-observable phenomenon. As a consequence, the readers can judge for themselves whether or not the main point of the thesis (10) holds. The same does not apply to the previous computational approaches simply because they do not provide an evaluation procedure. Their results often appear arbitrary, and cannot really be judged for this reason. The presented evaluation method is limited to matrix subject positions, and the accuracy is still not very high. But it can be extended to a wider range of utterance components as shown in Chapter 7, and other languages can be used for the same purpose. Thus, we can increase the coverage and the accuracy of the evaluation beyond what is presented here.

The thesis also covers a wider range of linguistic constructions, including various real-text properties, than previous work. Although the lexicon and the grammar still need to be extended, the information-structure analysis can be applied to a new set of realistic texts for further evaluation with little adjustment in terms of the theory of information structure. Thus, we have overcome Levinson's [1983] skepticism about the applicability of information-structure analysis for an arbitrarily complex linguistic structure.

There is one other factor associated with the main contribution. That is, the theory is made sufficiently explicit so that it is readily formalized and implemented as a procedure. This development contrasts with the situation where most theoretical works in linguistics are at a level that does not easily allow formalization and implementation. It also contrasts with most computational approaches, which lack the connection between their procedure and linguistic theories.

In addition to the above, the thesis contributes several points to the field of computational linguistics. By adopting a grammar-based parser, albeit one that is rather flexible in terms of dealing with constituency, the implementation of the theory retains the ability of precisely capturing various syntactic and semantic properties, and can integrate pragmatic factors in a straightforward manner. This provides a precise connection between utterance-level linguistic description and certain discourse-level concepts.

As a backbone of the system, we developed a practical parser for the CCG framework, overcoming the potential problem of spurious ambiguity. This point should remove the skepticism surrounding parsing CCG.

The thesis develops a comprehensive formalization and implementation of structured meanings. This not only captures the informational contrast present at every step of derivation, but also provides a platform for other properties including ‘contrast’ in a more general way than existing applications of structured meaning.

Finally, we provide an analysis of Japanese from the view point of a modern information-structure analysis. The functions of Japanese particles and long-distance fronting have been under discussion for a long time. Unfortunately, even the current literature does not fully reflect the recent advancement in studies of information structure and referential status. The current work updates this situation and provides materials useful for language-specific and cross-linguistic analyses. In addition, through the discussion on both English and Japanese in terms of information structure and contextual linking, we are able to relate certain underlying mechanisms of various pragmatic functions.

## **Future Directions**

One natural continuation of the present work is to integrate the information-structure analyzer with the applications discussed in the Introduction. For example, in most machine translation projects, a parser is already built in. While not all types of parsers can recognize constituents as flexibly as CCG parsers can, we may still use the derived linguistic structure and identify information structure based on the present approach. Then, the results can be used for prediction of particles in Japanese and word order in, e.g., Turkish.

Another application that I have a great interest is a Computer-Assisted Writing system, which can analyze text readability with respect to information structure. During the development of the present thesis, we seriously considered this project as an application domain and proposed a prototype (Section 2.4). A preliminary result on analyzing journal abstracts gives an impression that this application would make a noble, useful tool for writers. But the idea was not pursued for the present thesis because of the difficulty with evaluation. But I still consider this as an interesting long-term project.

The evaluation method proposed in the present work concentrates on the information-structure status of grammatical subjects. We may extend this to components other than subjects as briefly touched on in Chapter 7. We may also use other languages that marks information structure differently from the way it is done in Japanese. Since the linguistic marking of information structure in a single language by no means covers all the constructions, a multi-lingual analysis seems to be required for a more complete coverage.

Another direction is to use larger-scale parallel corpora available on the Internet. We have seen that the current accuracy of the prediction is at a level comparable to the individual variation (for unconstrained translation). Thus, using a larger number of texts written by different individuals may yield similar results without obtaining multiple translations.

As we mentioned in Section 2.1, there is a related problem of identifying definiteness in English encountered in an application such as Japanese-English machine translation. Our position is that the definiteness-identification problem is distinct from the Identification Problem for information structure. But there is a great deal of overlap. Both problems contain basically the same components: definiteness marking, contextual linking, and information structure. It is interesting to see how much the present theory can tell about the relation between the two, both shared and distinct elements.

The present thesis separates important areas of reference, inference, and discourse structure. Further exploration about the connection between information structure and these areas is a challenging but exciting future work.

Finally, the analysis of the present work may also apply to second-language education both in English and Japanese. A student of Japanese may learn certain concrete information about the use of particles, which is often perceived difficult or vague. A student of English may learn

the functions of various constructions in terms of a fairly small number of properties including contextual linking.