University of Pennsylvania
ScholarlyCommons

IRCS Technical Reports Series
Institute for Research in Cognitive Science

August 1998

Computer-Assisted Writing System: Improving Readability with Respect to Information Structure

Nobo Komagata
University of Pennsylvania, komagata@linc.cis.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/ircs_reports

Komagata, Nobo, "Computer-Assisted Writing System: Improving Readability with Respect to Information Structure" (1998). *IRCS Technical Reports Series*. 64.
http://repository.upenn.edu/ircs_reports/64

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-98-19.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/ircs_reports/64
For more information, please contact libraryrepository@pobox.upenn.edu.

Computer-Assisted Writing System: Improving Readability with Respect to Information Structure

Abstract
Text coherence and readability in English can be significantly affected by the organization of information in an utterance. To assist writers in this respect, we implement a Computer-Assisted Writing system focusing on 'information structure'. The main challenge in this task is identification of information structure in texts. This paper shows that this can be done by checking discourse status and linguistic marking in utterances.

Keywords
Information structure, text, discourse

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-98-19.

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/ircs_reports/64

In NLP+IA98, Moncton, NB, Canada, Aug. 1998, pages 200-204.

Computer-Assisted Writing System: Improving Readability with Respect to Information Structure

Nobo Komagata*
Computer and Information Science // Institute for Research in Cognitive Science
3401 Walnut Street, Suite 400A, Philadelphia, PA 19104-6228, USA
komagata@linc.cis.upenn.edu

Abstract

Text coherence and readability in English can be significantly affected by the organization of information in an utterance. To assist writers in this respect, we implement a Computer-Assisted Writing system focusing on 'information structure'. The main challenge in this task is identification of information structure in texts. This paper shows that this can be done by checking discourse status and linguistic marking in utterances.

Keywords: Information structure, text, discourse

1 Introduction

In a book on how to write a research paper, (Booth *et al.* 95) argue that, in order to improve readability, one should order information from 'old' to 'new' in each utterance. They advise the reader to write, e.g., (*b*) rather than (*a*) below.

(1) *a.* The mitral valve could be permanently damaged if the patient has mitral valve prolapse and develops endocarditis. Medication that controls infection will not halt this damage. Only surgery which repairs the defective valve will achieve that goal.

b. If the patient has mitral valve prolapse and develops endocarditis, the mitral valve could be permanently damaged. This damage will not be halted by medication that controls infection. That goal will be achieved only by surgery which repairs the defective valve.

This is not a kind of advice existing grammar checkers can offer, but can be overlooked by non-native and even native speakers of English.

Our focus here is implementation of a Computer-Assisted Writing system which can assist writers in this respect. The crucial point is how to analyze the organization of information. For our investigation, we adopt the notion of 'information structure' widely studied in linguistics, e.g., (Vallduví & Engdahl 94). Roughly, the idea is that components of an utterance exhibit different degrees of informativeness with respect to the context and that they are often linguistically marked. This point is essential for a NL generation task for, say, Turkish because the word order in Turkish is mainly determined by information structure (Hoffman 96). The same point applies to speech generation in English. Contextually-appropriate intonation cannot be generated without this information (Prevost & Steedman 93). But analysis of information structure in texts remains a difficult problem in theory and practice.

In this paper, we present an implementation of a Computer-Assisted Writing system and demonstrate that the information structure in medical abstracts can be identified and that the result can be used as advice for the writer.[1]

* I am grateful to Mark Steedman for his support and comments. I also thank Gehard Jäger, Ellen Prince, Michael Strube, Bonnie Webber, and the reviewers for their comments. The research was supported in part by NSF Grant Nos. IRI95-04372, STC-SBR-8920230, ARPA Grant No. N66001-94-C6043, and ARO Grant No. DAAH04-94-G04 26.

[1] A more extensive literature review and the details about this project can be found at http://www.cis.upenn.edu/~komagata/papers.html.

2 Information Structure

A preliminary view about information structure is that it is a binary division of an utterance where the 'referent' of one component (*theme*) is already in the 'discourse context' and the other (*rheme*) is not necessarily so.[2] For example, the following analysis is possible: "John has a house. [The house]$_{Theme}$ [looks exotic]$_{Rheme}$". But in "John has a house. [The door]$_{Theme}$ [looks exotic]$_{Rheme}$", we still want to consider *the door* as the theme even though it is not explicitly introduced in the preceding discourse. In this case, the definite determiner establishes a 'contextual link' equivalent to the discourse-old status (Prince 92, for discussion). I argue that this point must be incorporated in the characterization of information structure as follows:[3]

(2) **Information structure** of an utterance is a binary (semantic) division of an utterance into the following complementary components:

- **Theme**: The component which is discourse-old *or* signaled by linguistic marking.
- **Rheme**: The complement of the theme.

The current implementation focuses on definite/indefinite distinction as the linguistic marking.

(3)

Discourse status	Old	New	
Linguistic marking		Def.	Indef.
Contextual link	Yes		No

Other cases are also being studied. Our point here is that this characterization is good enough to identify information structure in the medical domain where the effect of inference does not in practice affect the analysis of information structure.

This approach contrasts with the previous computational approaches, which underestimate either the contextual effects (Kurohashi & Nagao 94; Hajičová *et al.* 95) or the role of linguistic structure (Hahn 95). In this respect, we follow (Hoffman 96), but are more flexible with respect to the semantic type of contextual link.

Now, let us see how the above definition can be applied to (1*a*) (discourse-old elements are indicated by $\boxed{}$, and linguistic marking of contextual link is indicated as $\underset{\textit{type of marking}}{\textit{phrase}}$):[4]

(4) *i.* The mitral valve could be permanently damaged if the patient has mitral valve prolapse and develops endocarditis.

ii. [Medication that controls infection will not halt]$_{Rheme}$ [$\underset{\text{def}}{\text{this}}$ $\boxed{\text{damage}}$]$_{Theme}$.

iii. [Only surgery which repairs $\underset{\text{def}}{\text{the}}$ defective $\boxed{\text{valve}}$ will achieve]$_{Rheme}$ [$\underset{\text{def}}{\text{that}}$ goal]$_{Theme}$.

In (*ii*), *this damage* is both discourse-old (identified with *damaged* in (*i*) through a derivational relation) and linguistically marked for a contextual link. Assuming the shown division of the utterance, we can identify the information structure of (*ii*). This rheme-theme order is against (Booth *et al.* 95)'s 'theme first' advice, cf. (1*b*). (*iii*) is similar except that *goal* is not discourse-old. But the linguistic marking forces the reader to infer an appropriate referent from the previous utterance.

Another crucial point in the above demonstration is that the informational divisions observed in (4*ii*,*iii*) do not correspond to traditional phrase structure divisions such as subject-VP. Assuming Montague-style semantics and considering the tight coupling between information structure and grammar inherent in our characterization (2), we choose a grammar which can deal with a more 'flexible' notion of constituency, i.e., Combinatory Categorial Grammar (CCG) (Steedman 91).[5]

3 Implementation

The current system is designed for the lexical entries and linguistic constructions found in 16 medical abstracts (105 sentences, approximately 700 lexical entries) from *The Physician and Sportsmedicine*. The average and maximum sentence length are 17 and 48, respectively.[6]

[2] Theme and rheme roughly correspond to rather overloaded terms 'topic' and 'focus', respectively.

[3] We separate the issues of reference resolution and inference, which can be integrated with the current approach.

[4] The discourse-initial utterance has a special status and is skipped here.

[5] Although the phrase *the defective valve* in (4*iii*) also signals a contextual link, it cannot be a theme because it is embedded in a relative clause and an appropriate utterance division is not available (even in CCG).

[6] Compound sentences are (manually) divided into simple sentences to focus on the point.

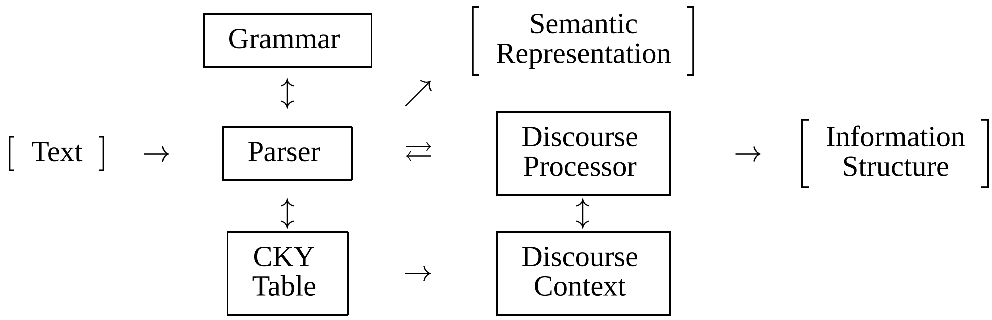

Figure 1: System Architecture

3.1 System Architecture

The system is implemented on a Sun Ultra E4000 2×250MHz Ultrasparcs with 320MB memory running SunOS 5.5.1. The code is written entirely in Sicstus Prolog Ver. 3. The program file is approximately 60KB and the grammar is about 100KB in size. The system architecture is shown in Fig. 1.

Since our grammar, CCG, can recognize flexible constituency in accordance with divisions of information structure, the discourse processing can proceed in parallel to parsing, represented by the bidirectional arrow '⇄' in Figure 1. Successful parses contained in the CKY table are stored and serve as a part of the discourse context.

The current scenario of Computer-Assisted Writing is that the writer inputs sentences one by one. The system analyzes the discourse status and the linguistic marking of each sentence, and reports the analyzed information structure.

3.2 Parser

To deal with flexible constituents and identify relevant linguistic marking, we adopt a CKY-style CCG parser (Komagata 97). One feature of our parser is that the CKY table contains *pointers* to categories, not categories themselves. This allows us to access the semantic representations already in the context without re-introducing a duplicate.

Now, the point about flexible constituency can be shown in the following example with two distinct derivations (with a simplified semantic representation for the NP).[7]

[7]The flexibility of CCG can result in multiple derivations of equivalent semantic representations. This does not poses a problem as long as redundant entries in a CKY table cell is eliminated.

(5)

	The patient	developed	endocarditis
a.	p_1	$\lambda x.\lambda y.d'(x)(y)$	e_1
		$\lambda y.d'(e_1)(y)$	
	$d'(e_1)(p_1)$		
b.	$\lambda f.f(p_1)$	$\lambda x.\lambda y.d'(x)(y)$	e_1
	$\lambda x.d'(x)(p_1)$		
	$d'(e_1)(p_1)$		

In (*b*), the semantic representation of *the patient* is 'type-raised' to make it a function as pursued in the Montagovian tradition. Then this type-raised function and the verb semantics can compose to derive a semantic representation corresponding to the subject-verb sequence. We skip the corresponding operations involving syntactic categories, but all of the above derivation stages are legitimate steps in CCG. The grammar can recognize the exact set of strings according to the lexical and rule specifications, with some freedom in terms of derivation.

The entire 16 abstracts can be parsed in about 5 minutes, or 3 seconds per sentence on average. The parser provides a practical platform with a capability to deal with flexible constituency corresponding to divisions of information structure.

3.3 Processing Information Structure

The process of identifying information structure proceeds bottom-up in parallel to the parsing process. At the local level, there are two tasks: analysis of discourse status and analysis of linguistic marking. As for discourse status, the system checks whether any semantic representation as a part of the current utterance is already in the discourse context. Consider the following short discourse:

```
Seg: the patient developed endocarditis . (5 words)

Result: cat(s(fin),develop-endocarditis-(def(the)-patient))

CPU time: 120 ms  Elapsed: 170 ms

*** IS-related info:
* def_marked(def(the)-patient)
* in_context(X^Y^(develop-X-Y))
* in_context(X^(develop-X-(def(the)-patient)))
* theme_rheme(X^(develop-X-(def(the)-patient)),endocarditis)
```

Figure 2: Sample Output

(6) *a.*

$$
\begin{array}{llll}
 & \text{What does} & \text{the patient} & \text{develop} \\
 & ?X & \lambda f.f\,(p_1) & \lambda x.\lambda y.d'\,(x)\,(y) \\
 & & \multicolumn{2}{c}{\boxed{\lambda x.d'\,(x)\,(p')}} \\
 & \multicolumn{3}{c}{d'\,(?X)\,(p_1)}
\end{array}
$$

b.

$$
\begin{array}{lll}
\text{The patient} & \text{developed} & \text{endocarditis} \\
\multicolumn{2}{c}{\boxed{\lambda x.d'\,(x)\,(p_1)}} &
\end{array}
$$

Once (*a*) is stored as a part of the context, the expression $\boxed{\lambda x.d'\,(x)\,(p_1)}$ in (*b*) can be identified with the equivalent expression in the context. Then the system sets up a link to the existing entry in the context and uses it for further processing. Thus there are no inherent limitations on what kind of semantic representation can be a contextual link. This property is not shared by the previous implementations.

The second local-level process is linguistic analysis. Since the current focus is on definite determiners, identification of NP is sufficient for this purpose. But the system analyzes the linguistic structure (albeit more flexibly than 'traditional' grammars), and can capture various grammatical conditions, cf. a text-based system (Hahn 95).

Next, the definition of information structure specifies a top-level process of identifying complementary components. This amounts to analysis of the two components at the last semantic composition. Again, due to the ability of CCG to recognize flexible constituents, various non-traditional divisions we have been observing can be captured this way.

A slightly simplified output of the program for (6*b*), where (6*a*) is assumed to be in the context, is shown in Fig. 2. A category here is a pair of a syntactic type, e.g., `S`, and a semantic representation, in the form of '`cat(Type,Sem)`'. `X^Y` represents $\lambda x.y$ in Prolog.

Next, the data can be classified into the following patterns with respect to information structure.[8]

(7) *a.* GOOD: Theme-rheme order in accordance with the 'theme first' preference.

b. BAD: Rheme-theme order against the preference.

c. UGLY: A sequence of all-rheme utterances gives an impression of a 'cut-and-paste' abstract.

Examples of GOOD and BAD are shown in Fig. 3. If we alter the information ordering in the utterances under consideration, e.g., by making a cleft (i.e., "what *VP* is *subject*") or switching NPs across the copula, the GOOD/BAD patterns appear to change. This way, we can informally evaluate the identification process of information structure.

Among 105 sentences in 16 abstracts, the system has identified 21 GOOD and 3 BAD cases, generally in accordance with our informal assessment. Now, the remaining question is whether the proposed theory can generalize to a larger set of abstracts. While we expect that the system works correctly for the case specified in this paper, the following possibilities also exist. First, identification of information structure may be incomplete due to incomplete specification of linguistic marking. Second, the discourse structure of the text and linguistically-unmarked inference may affect the identification process, but these aspects are clearly separated in the current formulation.

[8] We assume that there is no all-theme type sentence in the current domain.

- Abstract 10
  - *i.* (Title) Diagnosing Posterior Cruciate Ligament Injuries
  - *ii.* [$\boxed{\text{Posterior cruciate ligament (PCL) injuries}}$]$_{Theme}$ [are difficult to detect because patients rarely present with findings that suggest a severe ligament $\boxed{\text{injury}}$]$_{Rheme}$. (GOOD)
  - *iii.* (contd.)
- Abstract 7
  - *i.* (Title) Atypical Pneumonia in Active Patients: Clues, Causes, and Return to Play
  - *ii.* [$\boxed{\text{Atypical}}$ $\boxed{\text{pneumonias}}$ can affect young, otherwise healthy individuals who have close contact with one another, such as athletes in team sports]$_{Rheme}$.
  - *iii.* [Symptoms, which often progress gradually, may mimic an upper respiratory tract infection]$_{Rheme}$.
  - *iv.* [Mycoplasma, chlamydia, and legionella organisms, along with certain viruses, are]$_{Rheme}$ [$\underset{\text{def}}{\text{the}}$ usual $\boxed{\text{atypical}}$ $\boxed{\text{pneumonia}}$ agents]$_{Theme}$, (BAD)
  - *v.* (contd.)

Figure 3: GOOD and BAD Examples

4 Conclusion

This paper presents an implementation of a Computer-Assisted Writing system which can advise the writer of text readability with respect to information structure.

The future directions include integration of (i) a reference-resolution module partially involving user interaction, and (ii) a generation module to offer a contextually-appropriate alternative. The theory is also applicable to translation (Hoffman 96) and speech generation. For the latter, if utterance (4*iii*) is fed to the Bell Labs Text-to-Speech (TTS) (Lucent Technologies 97), *that goal* receives an incorrect pitch accent. With our theory, the TTS could deal with a wider range of texts in known domains.

References

(Booth *et al.* 95) Wayne C. Booth, Gregory G. Colomb, and Joseph M. Williams. *The Craft of Research*. University of Chicago Press, 1995.

(Hahn 95) Udo Hahn. 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. 1995.

(Hajičová *et al.* 95) Eva Hajičová, Hana Skoumalová, and Petr Sgall. An automatic procedure for topic-focus identification. *Computational Linguistics*, 21(1):81–94, 1995.

(Hoffman 96) Beryl Hoffman. Translating into free word order languages. In *COLING-96*, pages 556–561, 1996.

(Komagata 97) Nobo Komagata. Efficient parsing for CCGs with generalized type-raised categories. In *IWPT97*, pages 135–146, 1997.

(Kurohashi & Nagao 94) Sadao Kurohashi and Makoto Nagao. Automatic detection of discourse structure by checking surface information in sentences. In *COLING-94*, pages 1123–1127, 1994.

(Lucent Technologies 97) Lucent Technologies. Welcome to our multilingual text-to-speech systems (http://www.bell-labs.com/project/tts/), 1997.

(Prevost & Steedman 93) Scott Prevost and Mark Steedman. Generating contextually appropriate intonation. In *EACL6*, pages 332–340, 1993.

(Prince 92) Ellen F. Prince. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins, 1992.

(Steedman 91) Mark Steedman. Structure and intonation. *Language*, 67:260–296, 1991.

(Vallduví & Engdahl 94) Enric Vallduví and Elisabet Engdahl. Information packaging and grammar architecture. In *NELS25*, 1994.