A Paraphrase-Based Exploration of Cohesiveness Criteria

INUI Kentaro[†] [‡] and NOGAMI Masaru[†]
[†] Department of Artificial Intelligence, Kyushu Institute of Technology, JAPAN
[‡] PRESTO, Japan Science and Technology Corporation, JAPAN
{inui,m_nogami}@pluto.ai.kyutech.ac.jp

Abstract

This paper proposes an empirical approach to the development of a computational model for assessing texts according to cohesiveness. We argue that the NLG technologies for the generation of structural paraphrases can be used to efficiently create what we call a cohesion-variant parallel corpus, which would serve as a good resource for empirical acquisition of cohesiveness criteria. We also present our pilot case study, in which we took a particular type of paraphrasing that separates a relative clause from a sentence. We have so far created a cohesion-variant parallel corpus containing 499 cohesive instances and 841 incohesive instances. Based on this corpus, we conducted a preliminary experiment on cohesion evaluation, obtaining encouraging results.

1 Introduction

In NLP tasks such as translation, summarization and text generation, where a system produces *texts* as its output, the cohesiveness of output texts is an important criterion for assessing the system's performance. An output text should not be just a collection of syntactically and semantically well-formed sentences, but is required to be well-organized also at the discourse level. Cohesive relations (i.e. coreference relations, rhetorical relations, etc.) between entities contained in a text should be properly realized by means of linguistic cohesive devices. According to Halliday and Hasan (1976) and Halliday (1994), there are four types of cohesive devices in English: reference, ellipsis, conjunction, and lexical cohesion. A system thus needs to know how to use such cohesive devices effectively[1].

The translation and summarization research communities are, in fact, increasingly getting concerned with the notion of cohesion, though only recently, as technologies for intra-sentential processing make progress. For example, Marcu et al. (2000) proposes a computational model for transforming rhetorical structures between a source and target languages, aiming at the improvement of translation quality. Mani et al. (1999) proposes to incorporate the cohesion-concerned revision process into summarization.

What is commonly required in such tasks is a technology for assessing a given text according to cohesiveness. Such a technology would enable us to design, for example, a translation model which would choose the best cohesive translation from more than one generated candidate. Alternatively, it might also be feasible to design a framework where a system would revise an initial translation according to cohesion. In this paper,

[1] In this paper, we use the term *cohesiveness* to refer to the degree to which cohesive devices are properly chosen to realize cohesive relations in a text, whether it is quantified or not. It is a matter of surface realization or sentence planning (Wanner and Hovy, 1996; Reiter and Dale, 1999). Here we distinguish it from the notion of *coherence*, which is typically used to refer to the degree to which the text contents themselves are well-organized at the conceptual level, and thus a matter of content selection/organization. This paper concentrates its focus on the former stratum.

we argue that the NLG technologies for generating certain types of structural paraphrases can be used to efficiently create training and testing data which would serve as a good resource for empirical corpus-based study on cohesiveness evaluation.

To give our intuition to the reader, let us consider the following example, where (1t.1) and (1t.2) are both paraphrases of a source passage (1s)[2]:

(1s) Småland, which is located to the south-west of Stockholm, is called "The Kingdom of Glass". The reason is that there are sixteen glass manufacturers in this area.

(1t.1) Småland is located to the south-west of Stockholm. It is called "The Kingdom of Glass". The reason is that there are sixteen glass manufacturers in this area.

(1t.2) Småland is called "The Kingdom of Glass". It is located to the south-west of Stockholm. The reason is that there are sixteen glass manufacturers in this area.

One paraphrase (1t.1) is cohesive as well as the original passage (1s). The other paraphrase (1t.2) cannot be considered cohesive, however, since the REASON relation between the first and third sentences is interfered by the second sentence.

As suggested by this example, particular sorts of structural paraphrasing have the effect of changing some aspects of the textual structure of a given original text. Sentence division/aggregation, clause order scrambling, topicalization, extraposition[3], and voice-switching are typical sorts of such structural paraphrasing. These sorts of paraphrasing are likely to change either of discourse-relation scoping, theme-rheme chaining, coreference chaining, salience, and so on. This change may preserve the original cohesiveness, but may also be likely to break it. Making use of this nature, one can systematically and efficiently create a large text collection containing both cohesive and incohesive instances in parallel, which would then serve as a resource for exploration of cohesiveness criteria. We call such a text collection a *cohesion-variant parallel corpus*.

In this paper, we concentrate our focus only on the cohesiveness of the linguistic realization level rather than on the coherence of the conceptual level. Our present goal is, therefore, to create a computational model that can evaluate and compare a given set of paraphrase variants associated with the same contents according to local cohesiveness. A pair of passages (1t.1) and (1t.2) above is an example of such a paraphrase variant set. The scope of the variations of linguistic cohesive devices we would like to consider here largely covers the matters of sentence planning (or micro planning) (Wanner and Hovy, 1996; Reiter and Dale, 1999) including referring expressions, discourse markers, sentence grouping, clause configuration, topicalization, etc.

In the following sections, we first review the literature on choice of cohesive devices in text generation in section 2. We next describe an overview of our paraphrase-based approach to this issue in section 3. We then present our pilot case study, in which we took a particular type of paraphrasing that separates a relative clause from a sentence, reporting the results of a preliminary experiment, in section 4 and 5.

2 Choice of cohesive devices in text generation

One may expect that criteria or techniques for cohesiveness evaluation should be easily found in the literature on text generation for a couple of reasons:

- There have been quite a few cohesion-related works in this field. For example, works on discourse marker choice (Scott and de Souza, 1990; Vander Linden, 1994; Grote and Stede, 1998; Oates, 1999), generation of referring expressions (Dale, 1992), and clause aggregation based on discourse relation (Dalianis and Hovy, 1996; Show, 1998) are all related to linguistic realization of cohesive relations.

- Text generation has the advantage of serving as a good device for systematically producing diverse text variants from the same input. Imagine, for example, that a NLG sys-

[2] This example is a congruent translation of a passage originally written in Japanese.

[3] A typical example is the transfer from a simple-clause sentence to a cleft sentence.

tem could generate passages (1s), (1t.1) and (1t.2) in parallel from the same content specifications. If given such a collection of text variants containing both cohesive and incohesive instances in parallel, one could carry out steady explorations of cohesiveness criteria, since having negative instances besides positive ones would significantly facilitate generalization.

Unfortunately, however, there have so far been very few works in this field that have shown comprehensive and concrete criteria of cohesiveness. This is supposed to be partly due to the following problems:

- In text generation, input topics tend to strongly depend on each application system, which makes it difficult to conduct large-scale experiments to generate text variants of diverse contents. This seems to have prevented transfer of cohesion-related technologies in text generation to other fields such as translation and summarization that are required to handle relatively unrestricted texts.

- Ideally, for computational purposes, cohesiveness criteria should be represented declaratively as, for example, a set of declarative constraints, so that they could be reemployed in other NLP tasks. In most existing generation systems, however, the process of considering cohesiveness tends to be distributed over a number of choice points in the search space of generation (more specifically, for example, decision experts in blackboard-based a sentence planner (Wanner and Hovy, 1996; Grote and Stede, 1998)), which has obstructed reuse of implemented knowledge.

- While NL understanding/analysis communities have recently developed empirical corpus-based approaches gaining remarkable success, it seems that the NLG community has not made so much effort on this frontier. For example, it has not yet accumulated significant *shared* text data (corpora) that are richly annotated for the purpose of empirical discourse analysis. Coherence-oriented knowledge proposed so far in the literature, e.g. the assumptions and heuristics proposed by Scott and de Souza (1990), would be more easy to refine, extend or customize, and thus would be more reusable, if it were available together with the text data from which it had been extracted.

Concerning the first and third problems, one may see a remarkable exception in Marcu's empirical approach (Marcu, 1997). Marcu first automatically sampled a large collection of text fragments including discourse markers in question, then carried out manual annotation, and finally attempted to acquire local constraints based on statistics. Marcu's approach is considered fairly powerful, yet it still has a weakness: Since it is not designed to use negative instances in the constraint acquisition, it requires huge amount of training data, which then requires considerable cost for manual annotation.

3 Paraphrase-based generation of cohesion-variant parallel corpora

Our approach can be decomposed into the following steps:

1. **Manual production of paraphrases:** Given a collection of several-clause-long passages, we first manually create a structural paraphrase for each passage so that it preserves the cohesiveness of the original passage. Each paraphrase created in this step is called a *core positive instance*.

2. **Manual extraction of a choice system:** We next manually analyze the differences between the core positive instances and their original passages to extract a choice system of paraphrasing that exhaustively covers all the core positive instances. Here, we assume a choice system to be a sort of a system network in the systemic sense (Halliday, 1994).

3. **Implementation:** We implement the choice system on a paraphrase generator.

4. **Automatic generation of annotated paraphrases:** We then generate diverse paraphrases from each of given input passages systematically by making random choices on

the choice system. In this process, the generator simultaneously annotates each paraphrase instance with rich tags to indicate all the choices made to generate it as well as various syntactic and semantic attributes of it.

5. **Manual cohesiveness evaluation:** Theoretically, the resultant set of paraphrases should cover all the core positive instances. In addition, it also include a much larger number of other newly generated paraphrases. For each instance of the latter, we manually judge whether it is positive (i.e. cohesive) or negative (i.e. incohesive), and annotate it accordingly.

As a result of steps 1 through 5, we obtain an annotated cohesion-variant parallel corpus. It consists of *paraphrase groups*, each of which further consists of both cohesive and incohesive paraphrases associated with the same clause set.

6. **Modeling and testing of cohesiveness criteria:** Such a sufficiently large annotated corpus being obtained, one should be able to employ it to create and test computational models for cohesiveness evaluation. For this modeling step, one might also be able to apply some recently advanced machine learning techniques, since the task of cohesiveness evaluation we consider here can also be simply regarded as a classification problem where a given passage (paraphrase) is to be classified into two classes: cohesive or incohesive.

Our paraphrase-based approach has the following advantages, while preserving the advantage of previous verification-by-generation approaches.

- Structural paraphrasing tends to change only a very small part of a given original sentence. This makes it easier to implement a structural paraphrase generator that guarantees at least the intra-clausal syntactic/semantic well-formedness of its output, compared with the case of generation from knowledge base.

- It is also relatively easy to generate text variants of sufficiently diverse contents, since paraphrasing does not require either application-dependent artifactual input or a grammar and lexicon that fully cover the texts to generate.

- Manual semantic/discourse annotation is required in principle only for source instances; a much larger number of derivative instances can be annotated fully automatically. This facilitates scaling up of an instance collection.

- The second advantage potentially enables us to make so-called *selective sampling*, which has been empirically proven to effectively accelerate learning, while reducing manual annotation costs, in many knowledge acquisition tasks, e.g. (Fujii *et al.*, 1998).

4 A case study

We conducted a pilot case study, taking a particular type of paraphrasing which separates a relative clause from a given sentence as in example (1) in section 1. Hereafter, for simplicity, we use the term *paraphras*{*e,ing*} to refer to paraphras{e,ing} of this type, as far as the present case study is concerned. Furthermore, for convenience, we call the sentence originating from a relative clause a *satellite sentence* (or simply a *satellite*), and the sentence that consists of the remaining constituents of the source sentence a *nucleus sentence* (or simply a *nucleus*).

The target language we have so far explored is only Japanese. Note, however, that our methodology is expected to be in principle equally applicable to any language. Example (2) below is an actual example of Japanese paraphrasing:

(2s) [*Sweden-no* (Sweden-POS) *shuto* (capital-APPOS) *Stockholm-no* (of Stockholm) *nansêbu-ni* (to the south-west) *itisuru* (to be located-ADNOM)]$_{REL\_CLS}$ *Småland-tihô-wa* (Småland-TOP) *betumê* (another name) *“garasu-no ôkoku”-to* (as “Kingdom of Glass”) *yobareteiru* (to be called).

(2t) ⟨satellite⟩ *Småland-tihô-wa* (Småland-TOP) *Sweden-no* (Sweden-POS) *shuto* (capital-APPOS) *Stockholm-no* (of Stockholm) *nansêbu-ni* (to the south-west) *itisuru* (to be located).
⟨nucleus⟩ *kono-tihô-wa* (this region-TOP) *betumê* (another name) *“garasu-no ôkoku”-to* (as “Kingdom of Glass”) *yobareteiru* (to be called).

Japanese relative clauses can be classified as either gapping or non-gapping. While gapping relative clauses contain a unique gap for the modified head, the associated case slot of which can be a complement or adjunct, non-gapping relative clauses do not contain any gap. The former class of relative clauses can be further semantically classified into restrictive or non-restrictive relative clauses. Among those three subtypes, for our present study, we restricted the source objects of paraphrasing to non-restrictive gapping relative clauses, since this type of relative clauses can be separated from the matrix clause most straightforwardly.

We first collected 275 non-restrictive gapping relative clauses from newspaper articles of diverse genres (1,840 sentences in total) excerpted from the Kyoto corpus (Kurohashi and Nagao, 1997). For each of them, we manually created a core positive instance, taking its context into account. We next manually analyzed those instances, and obtained a choice system consisting of seven major simultaneous choice points as follows:

(c1) **Tense and aspect:** whether the tense of the satellite should be of the *ta* (past) form or *ru* (non-past) form, and whether the aspect should be of the *teiru* (progressive/resultative) form or *ru* (base) form

(c2) **Case marker alteration:** the case marker *no* used as a nominative case marker in a relative clause should be obligatorily replaced with the proper subjective case marker *ga* in the satellite sentence

(c3) **Punctuation:** punctuation should be changed accordingly

(c4) **Connective:** whether the rhetorical relation between the nucleus and satellite should be verbalized as a connective expression or not, and which expression should be chosen if necessary

(c5) **Sentence order:** whether the nucleus sentence precedes or the satellite sentence precedes

(c6) **Topicalization:** whether the filler of the gap of the satellite should be topicalized or not

(c7) **Copulativization:** whether the satellite should be further transferred to construct a copula or not:

(3s) [*NTT-ga* (NTT-NOM) *4-gatu kara* (from April) *têkyô-suru* (to provide-ADNOM)]$_{REL\_CLS}$ *zisedai-kôsoku-tûsin-kaisen* (new generation telecommunication network)
(the new generation telecommunication network, which NTT will provide from April)

(3t.1) ⟨non-copula⟩ *NTT-ga* (NTT-NOM) *4-gatu kara* (from April) *zisedai-kôsoku-tûsin-kaisen-o* (new generation telecommunication network-ACC) *têkyô-suru* (to provide).
(NTT will provide a new generation telecommunication network from April.)

(3t.2) ⟨copula⟩ *zisedai-kôsoku-tûsin-kaisen-wa* (new generation telecommunication network-TOP) [*NTT-ga* (NTT-NOM) *4-gatu kara* (from April) *têkyô-suru* (to provide)]$_{REL\_CLS}$ *sâbisu-da* (to be a service).
(The new generation telecommunication network is a service that NTT will provide from April.)

(c7) **Anaphora/ellipsis:** Each anaphoric expression and ellipsis should be reconsidered with the options including at least the following:

- NP with a demonstrative adjective "*kono*/*sono* (this/that)"
- bare NP without a demonstrative adjective
- head noun with a demonstrative adjective
- demonstrative pronoun "*kore*/*sore*/ (this/that)"
- personal pronoun
- ellipsis (zero pronoun)

We then implemented the above choice system on our paraphrasing engine FUNE (Fujita *et al.*, 2000), and obtained 1,343 paraphrase instances from the 195 source instances, which were those randomly sampled from the above 275 source instances. To generate these paraphrases, we made random choices only for the choice points (c4) to (c7), while making an optimal choice for each of the rest, (c1) to (c3), since our preliminary investigation have proven the latter set of choice points to be almost independent of the context.

For the input to FUNE, we provided the following sorts of information:

- morphological and dependency structure information (given by the Kyoto corpus)

- semantic information (semiautomatically annotated) such as the grammatical role of the gap of a relative clause

- textual information (semiautomatically annotated) such as the rhetorical relations between clauses (Mann and Thompson, 1987) and the antecedent of each anaphor/ellipsis

Here, we mean by “semiautomatically annotate” that the preprocessing module analyzed the input to obtain semantic/textual information while leaving uncertain parts of analysis in our hands.

Finally, we manually assessed all the paraphrase instances. 449 instances were judged to be acceptably cohesive (positive), 841 instances unacceptable (negative), and 53 instances were left unjudged. When more than one positive instance were derived from a single source instance, we further ranked them.

The assessment was carried out by two of us. Unfortunately, we were not able to estimate the agreement rate between the two assessors, since we had frequently discussed all the cases of which either of us had felt unsure. The psychological estimation of the feasibility of human judgment in this task will be a future work.

5 A cohesiveness evaluation model

The cohesiveness criteria can be modelled as a set of *constraints* and *preferences*. The constraints would discriminate between positive instances and negative instances, whereas the preferences would rank positive instances according to fluency. As mentioned before, to create such a computational model, one might be able to employ various machine learning techniques for classification problems. We considered, however, that for the present case study, which is still at the very preliminary stage, it should be more important to get a sense of the properties of the task by manual analysis. In this section, we briefly but exhaustively enumerate the hypothetical constraints and preferences we have so far obtained by manual analysis.

5.1 Clause ordering

We considered that one way to approach the issue of clause ordering would be to start with Minami’s linguistic theory on intra-sentential hierarchical structure (Minami, 1974). According to Minami, a Japanese sentence has a center-embedding hierarchical structure as illustrated in Figure 1 (a), where the event description level (A) is embedded in the speaker’s attitude level (B), which is then embedded in the presentation level (C), but not vise versa. Given this view of sentence structure, one can predict, for example, that a subordinate adverbial clause stating REASON (level B) can be embedded in another subordinate clause stating CONCESSION (level C), but not vice versa, as illustrated in Figure 1 (b). This constraint has, in fact, been used by NLP researchers such as Shirai et al. (1995) for disambiguation of intra-sentential inter-clausal dependency structures.

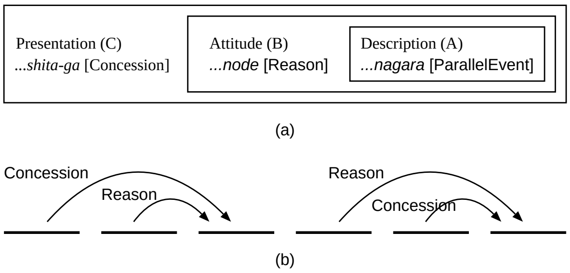

Figure 1: Hierarchical structure of Japanese sentence

In order to apply this constraint to our task of clause ordering, we need to extend it in the following respects:

- Since Minami’s theory covers only *intra*-sentential rhetorical structures, we need to prove whether it holds beyond sentence boundaries, and also whether it holds even if the nucleus precedes the satellite (In a Japanese sentence, a subordinate (satellite) clause always precedes the matrix clause (nucleus) it depends on.)

- The ELABORATION relation, which is one of the most common rhetorical relations that appear in inter-sentential rhetorical structures, is out of scope in Minami’s theory since the ELABORATION relation does not appear in adverbial inter-clausal dependencies. We need to investigate where the ELABORATION relation should be located in Minami’s hierarchy of rhetorical relations.

Our analysis has so far supported the following constraints and several well-known preferences, although obviously they still need further investigation and refinement.

Constraint 1.1 If three continuous discourse segments constitute either of the rhetorical patterns (A) or (B) shown in Figure 2, relation R_1 should be of a higher level of the rhetorical hierarchy than relation R_2, where:

- ELABORATION constitutes a new class whose level in the hierarchy is higher than the level (B) (e.g. REASON), and lower than the level (C) (e.g. CONCESSION), and
- the constraint holds beyond sentence boundaries, and is independent of the order of the nucleus and satellite, except that pattern (A) is not acceptable if R_2 is ELABORATION.

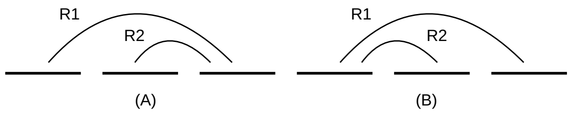

Figure 2: Local patterns of rhetorical dependency structure

The *Småland* example taken in section 1 is a good example that satisfies this constraint, where (4t.1) has no embedding, (4t.2) is the case where R_1=ELABORATION and R_2=REASON, and (4t.3) is the case where R_1=REASON and R_2=ELABORATION:

(4t.1) *Småland-tihô-wa* (Småland-TOP) *Sweden-no* (Sweden-POS) *shuto* (capital-APPOS) *Stockholm-no* (of Stockholm) *nan- sêbu-ni* (to the south-west) *itisuru* (to be located).
kono-tihô-wa (this region-TOP) *betumê* (another name) *"garasu-no ôkoku"-to* (as "Kingdom of Glass") *yobareteiru* (to be called).
16-mo-no (sixteen-EMPHASIS) *garasu-kôjô-ga* (glass manufacturers-MON) *kono-tihô-ni* (in this region) *tenzai-site-iru-kara-da* (to exist-REASON).
(Småland is located to the south-west of Stockholm, the capital of Sweden. It is also called "Kingdom of Glass". The reason is that there are sixteen glass manufacturers in this area.)

(4t.2) *Småland-tihô-wa* (Småland-TOP) *betumê* (another name) *"garasu-no ôkoku"-to* (as "Kingdom of Glass") *yobareteiru* (to be called).
16-mo-no (sixteen-EMPHASIS) *garasu-kôjô-ga* (glass manufacturers-MON) *kono-tihô-ni* (in this region) *tenzai-site-iru-kara-da* (to exist-REASON).
Småland-tihô-wa (this region-TOP) *Sweden-no* (Sweden-POS) *shuto* (capital-APPOS) *Stockholm-no* (of Stockholm) *nan- sêbu-ni* (to the south-west) *itisuru* (to be located).
(Småland is called "Kingdom of Glass". The reason is that there are sixteen glass manufacturers in this area. Småland is located to the south-west of Stockholm, the capital of Sweden.)

(4t.3) ∗ *Småland-tihô-wa* (Småland-TOP) *betumê* (another name) *"garasu-no ôkoku"-to* (as "Kingdom of Glass") *yobareteiru* (to be called).
kono-tihô-wa (this region-TOP) *Sweden-no* (Sweden-POS) *shuto* (capital-APPOS) *Stockholm-no* (of Stockholm) *nan- sêbu-ni* (to the south-west) *itisuru* (to be located).
16-mo-no (sixteen-EMPHASIS) *garasu-kôjô-ga* (glass manufacturers-MON) *kono-tihô-ni* (in this region) *tenzai-site-iru-kara-da* (to exist-REASON).
(∗ Småland is called "Kingdom of Glass". It is located to the south-west of Stockholm, the capital of Sweden. The reason is that there are sixteen glass manufacturers in this area.)

Constraint 1.2 The satellite of an ELABORATION relation cannot precede the nucleus, except for the case where the satellite has no preceding context.

Preference 1.1 If there is a coreference relation between two segments, they are preferred to be adjacent to each other.

Preference 1.2 If there is a temporal SUBSEQUENCE relation between two segments, they are preferred to be placed in that temporal order.

Preference 1.3 If the nucleus and satellite sentences are in the CONTRAST relation, the former is preferred to precede.

5.2 Discourse markers

Concerning discourse markers, we have not widely explored options for either of marker occurrence, marker placement and marker selection. For the moment, we have implemented only the following constraint for the experiment we will describe in the next section.

Constraint 3.1 The rhetorical relation should be verbalized by means of a proper connective expression in those cases including the following:

- a case where the rhetorical relation is REASON, and the satellite follows the nucleus
- a case where the rhetorical relation is CONCESSION, and the nucleus follows the satellite

5.3 Topicalization

Constraint 2.1 If the gap of the relative clause is associated with the nominative case, the gap filler should be topicalized in the satellite sentence, except for the case where the satellite is of the form “. . . *kara-da* (it is because . . .)” or the satellite is placed at the head of the text.

s. [*Prodigy-de* (on Prodigy) *sâbisu-o* (service-ACC) *okonatte-iru* (to be providing-ADNOM)]$_{REL\_CLS}$ *Access Atlanta-no* (of Access Atlanta) *kihon-ryôkin-wa* (basic rate-TOP) *getugaku* (monthly) *yaku-7-doru* (around 7 dollars-COPULA).
(The basic rate of Access Atlanta, which provides services on Prodigy, is around 7 dollars.)

t. *Access Atlanta-no* (of Access Atlanta) *kihon-ryôkin-wa* (basic rate-TOP) *getugaku* (monthly) *yaku-7-doru* (around 7 dollars-COPULA).
(The basic rate of Access Atlanta is around 7 dollars.)
<u>*Access Atlanta-wa*</u> (Access Atlanta-TOP) *Prodigy-de* (on Prodigy) *sâbisu-o* (service-ACC) *okonatte-iru* (to be providing-ADNOM).
(Access Atlanta provides services on Prodigy.)

Preference 2.1 If the gap of the relative clause is associated with the nominative case and the satellite is of the form “. . . *kara-da* (it is because . . .)”, then the gap filler is preferred not to be topicalized (Kuno, 1974).

Preference 2.2 If the gap of the relative clause is not associated with the nominative case, the gap filler is preferred to be topicalized.

5.4 Anaphora and ellipsis

Several works have explored the relation between rhetorical structure and reference in English (Fox, 1987; Cristea *et al.*, 2000; Grosz and Sidner, 1986; Grosz *et al.*, 1995). Japanese reference, on the other hand, has been studied from a different perspective, being associated mainly with the linear nature of texts as in the centering theory (Kameyama, 1986; Walker *et al.*, 1994). Considering that choice of referring expressions is in itself a quite large issue, we have been exploring it separately from this paraphrase-based exploration (Hashimoto, 2001). We will not go into the detail here, since we have so far implemented only the following well-known constraint and preference for the experiment.

Constraint 4.1 If two neighboring sentences have different topics (themes), the topic of the following sentence should not be omitted.

Preference 4.1 If two neighboring sentences share the same topic (theme), the topic of the following sentence is preferred to be omitted.

6 A preliminary experiment

We conducted a preliminary experiment to test the hypothetical model described in the last section. For the test set, we used 133 positive and 227 negative derivative instances associated with 100 source instances that were randomly sampled from the training set we used for the model development. This experiment was thus a closed test (For open testing, we are currently planning to create a new large-scale test set by employing several subjects). We then implemented the cohesiveness evaluation module, which was designed to apply the above-mentioned constraints and preferences to a given paraphrase group.

For the performance evaluation, we first tested the validity of the constraints by investigating how correctly they can discriminate between positive and negative instances. The result is shown in Table 1, where the recall is the ratio of the instances that the system correctly judged positive (116 instances) to all the positive instances (133 instances), whereas the precision is the ratio of the instances that the system correctly judged positive (116 instances) to all the instances that the system judged positive (155 instances).

Our error analysis revealed that, among the 39 cases where the system missed rejecting an incohesive instance, 16 cases were simply due to the inadequacy of knowledge about copulativization and anaphora/ellipsis, which we have not fully treated yet and will cope with in the next research step. Taking this into account, although our experiment is so far a closed test, the result can be considered to prove that our paraphrase-based empirical approach is reasonably promising, or at worst worth proceeding further.

Next we evaluated the performance of the preferences for ranking the positive instances. The result is shown in Table 2, where we count only the cases where a paraphrase group had more than one positive instance that satisfy all the constraints. “*complete*” denotes the cases where the ranking given by the system completely agreed with the ranking by the human judges, whereas “*best*” denotes the cases where the system agreed with the human at least on the best-ranked instance. This result is also encouraging, although the scale is still too small to support any statistically verified conclusion.

Table 1: The discrimination performance of the constraints

	human judgment		total
	positive	negative	
system positive	116	39	155
system negative	17	188	205
recall	87.2%		
precision	74.8%		

Table 2: The ranking performance of the preferences

	agreed	disagrd	total	accuracy
complete	26	11	37	70.3%
best	35	2	37	94.6%

7 Conclusion

We argued that the NLG technologies for the generation of structural paraphrases can be used to create cohesion-variant parallel corpora at reasonable cost. Such corpora would contain diverse text variants associated with the same contents, some of which are cohesive (i.e. positive), while some of which are not (i.e. negative). Having negative instances besides positive ones is expected to facilitate empirical acquisition of declarative and thus reusable constraints and preferences on local cohesiveness.

We also described our pilot case study, in which we adopted the particular type of paraphrasing that separates a relative clause from a sentence, and reported the results of a preliminary experiment. We then reported our preliminary experiment. The results we have so far obtained seem to us encouraging. The scale of the experiment is still too small to derive any statistically verified conclusion. To the best of our knowledge, however, there have been very few works in the NLG literature which had presented any experiment on cohesive device choice of even such a scale.

Aiming at the substantial scaling up of this study, we are currently developing a more sophisticated computational environment for paraphrasing and tagging to maximally reduce the manual cost. It is designed also to realize other diverse types of paraphrasing. Once a sufficiently large cohesion-variant parallel corpus is obtained, it will be highly worthwhile to apply machine learning techniques. The need to conduct open tests is also obvious.

A sufficiently comprehensive model for cohesiveness evaluation being acquired, it would be applicable to various NLP tasks such as translation, summarization, and text simplification. Our work will also be directed to the incorporation of such a model into the Japanese text simplification system we are currently developing, which is designed to assist congenitally deaf readers (Inui, 2001).

Acknowledgments

The authors would like to thank Dr. Michael Zock (LIMSI) for his suggestive comments on an early draft of this paper. They would also like to thank the reviewers for their helpful and encouraging comments.

References

Cristea, D., Ide, N., Marcu, D., and Tablan, V. An empirical investigation of the relation between discourse structure and co-reference. In *Proceedings of the International Conference on Computational Linguistics*, 2000.

Dale, R. *Generating Referring Expressions*. The MIT Press, 1992.

Dalianis, H. and Hovy, E. Aggregation in natural language generation. In *Proceedings of the 7th Conference of the European Chapter of the Association for Cumputational Linguistics*, 1996.

Fox, B. *Discourse Structure and Anaphora*. Cambridge Studies in Linguistics, Cambridge University Press, 1987.

Fujii, A., Inui, K., Tokunaga, T. and Tanaka, H. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4), 1998.

Fujita, A., Inui, K., and Inui, H. An environment for constructing nominal-paraphrase corpora. In *Proceedings of the Regular Meeting of IEICE Technical Group on Thought and Language*, IEICE-TL2000-32, 2000. (In Japanese)

Grosz, B. J. and Sidner, C, L. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

Grosz, B. J., Joshi, A. K., and Weistein, S. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), 1995.

Grote, B. and Stede, M. Discourse marker choice in sentence planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, 1998.

Halliday, M. A. K. *An Introduction to Functional Grammar (Second Edition)*. Edward Arnold, 1994.

Halliday, M. A. K. and Hasan, R. *Cohesion in English*. Longman, 1976.

Hashimoto, S., Inui, K., Shirai, K., Tokunaga, T. and Tanaka, H. Generation of anaphoric expressions in Japanese sentence generation. *Proceedings of the Regular Meeting of IPSJ Special Interest Group on Natural Language Processing*, IPSJ-SIGNL-143, 2001. (In Japanese)

Inui, K. A text simplification system for congenitally deaf readers. In *Proceedings of the ANLP-2001 Workshop on Automated Paraphrasing*, 2001. (In Japanese. The English version will soon become available as a technical report.)

Kameyama, M. A property-sharing constraint in centering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1986.

Kuno, S. *The Structure of the Japanese Language*. MIT Press, 1974.

Kurohashi, S. and Nagao, M. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS)*, 1997.

Mani, I., Gates, B., and Bloedorn, E. Improving Summaries by Revising Them. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1999.

Mann, W. C. and Thompson, S. A. Rhetorical structure theory: A theory of text organization. In Polany, L. ed. *Discourse Structure*, Ablex, Norwood, 1987.

Marcu, D. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 1997.

Marcu, D., Carlson, L., and Watanabe, M. The automatic translation of discourse structures. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2000.

McDonald, D. D. Natural language generation. In Dale, R. et al eds, *Handbook of Natural Language Processing*, Marcel Dekker, 2000.

Minami, F. *Gendai Nihongo no kôzô* (*The Structure of Contemporary Japanese*). Taishûkan Shoseki (Taishûkan Press), 1974. (In Japanese)

Oates, S. L. State of the art report on discourse markers and relations. Technical report, ITRI-99-08, University of Brighton, 1999.

Reiter, E. and Dale, R. *Building Natural Language Generation Systems*. Cambridge University Press, 1999.

Scott, D. and de Souza, C. Getting the message across in RST-based text generation. In Dale, R., Mellish, C. and Zock, M. (Eds.), *Current Research in Natural Language Generation*, Academic Press, 1990.

Shirai, S., Ikehara, S., Yokoo, A., and Kimura, J. A new dependency analysis method based on semantically embedded sentence structures and its performance on Japanese subordinate clauses. *Journal of Information Processing Society of Japan*, 36(10), 1995. (In Japanese)

Show, J. Clause aggregation using linguistic knowledge. In *Proceedings of the International Workshop on Natural Language Generation*, 1998.

Walker, M. A., Iida, M. and Cote, S. Japanese discourse and the process of centering. *Computational Linguistics*, 20, 1994.

Vander Linden, K. Generating precondition expressions in instructional text. In *Proceedings of the 15th Conference on Computational Linguistics*, 1994.

Wanner, L. and Hovy, E. The HealthDoc sentence planner. In *Proceedings of the 8th International Workshop on Natural Language Generation*, 1996.