An Empirical Study in Multilingual Natural Language Generation: What Should A Text Planner Do?

Daniel Marcu
Information Sciences Institute and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

Lynn Carlson
U.S. Department of Defense
Ft. Meade, MD 20755
lmcarls@afterlife.ncsc.mil

Maki Watanabe
Department of Linguistics
University of Southern California
Los Angeles, CA 90089
mwatanab@usc.edu

Abstract

We present discourse annotation work aimed at constructing a parallel corpus of Rhetorical Structure trees for a collection of Japanese texts and their corresponding English translations. We discuss implications of our empirical findings for the task of text planning in the context of implementing multilingual natural language generation systems.

1 Introduction

The natural language generation community has emphasized for a number of years the strengths of multilingual generation (MGEN) systems (Iordanskaja et al., 1992; Rösner and Stede, 1992; Reiter and Mellish, 1993; Goldberg et al., 1994; Paris et al., 1995; Power and Scott, 1998). These strengths concern the reuse of knowledge, the support for early drafts in several languages, the support for maintaining consistency when making changes, the support for producing alternative formulations, and the potential for producing higher quality outputs than machine translation. (The weaknesses concern the high-cost of building large, language-independent knowledge bases, and the difficulty of producing high-quality, broad-coverage generation algorithms.)

From an economic perspective, the more a system can rely on language independent modules for the purpose of multilingual generation, the better. If an MGEN system needs to develop language dependent knowledge bases, and language dependent algorithms for content selection, text planning, and sentence planning, it is difficult to justify its economic viability. However, if most of these components are language independent and/or much of the code can be re-used, an MGEN system becomes a viable option.

Many of the early implementations of MGEN systems have adopted the perspective that text planners can be implemented as language-independent modules (Iordanskaja et al., 1992; Goldberg et al., 1994), possibly followed by a *linearization* stage, in which discourse trees are re-written to reflect language-specific constraints (Rösner and Stede, 1992; Stede, 1999). Although such an approach may be adequate for highly restricted text genres, such as weather forecasts, it usually poses problems for less restricted genres. Studies of instruction manuals (Rösner and Stede, 1992; Delin et al., 1994; Delin et al., 1996) suggest that there are variations with respect to the way high-level communicative goals are realized across languages. For example, Delin et al. (1994) noticed that sentences (1), (2), and (3), which were taken from a trilingual instruction manual for a step-aerobics machine, yield non-isomorphic Rhetorical Structure (Mann and Thompson, 1988) analyses in English, French, and German respectively (see Figure 1).

English: [The stepping load can be altered^1] [by loosening the locking lever^2] [and changing the position of the cylinder foot^3]. (1)

French: [Pour modifier la charge d'appui,^{1}] [desserrer les levieres^2] [puis déplacer le pied des vérins^3] ([To modify the load stepping^1] [loosen the levers^2] [then change the foot of the cylinder foot.^{3}]) (2)

German: [Nach Lockern der Klemmhebel^2] [kann^1] [durch Verschieben des Zylinderfußes^3] [die Tretbelastung verändert werden.^{1}] ([After loosening of the levers^2] [can^1] [by pushing of the cylinder foot^3] [the load changed be,^{1}]) (3)

However, previous discourse studies do not estimate how ubiquitous such non-isomorphic analyses are. Are the examples above an exception or the norm? Are non-isomorphic analyses specific to discourse structures built across elementary discourse units of single sentences, or do they also occur across sentences and paragraphs? If non-isomorphism is ubiquitous, how should an MGEN system be designed in order to effectively deal with non-isomorphic discourse structures when mapping knowledge bases into multiple languages?

In this paper, we describe an experiment that was designed to answer these questions. To investigate

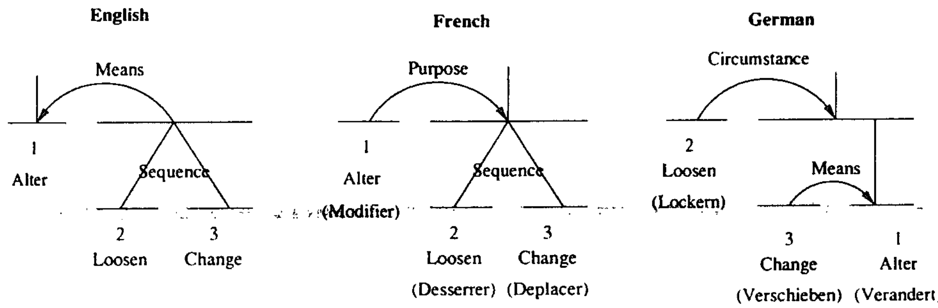

Figure 1: Contrasting multilingual discourse structure representations (Delin et al., 1994, p. 63)

how discourse structures differ across languages, we manually built a parallel corpus of discourse trees of newspaper Japanese texts and their corresponding English translations. In section 2, we present some of the problems specific to the construction of such a corpus. In section 3, we present our experiment and discuss our empirical findings. In Section 4, we discuss the implications of our work for the task of text planning, in the context of multilingual natural language generation.

2 Towards building a parallel corpus of discourse trees: an example

Consider, for example, Japanese sentence (4), a word-by-word "gloss" of it (5), and a two-sentence translation of it that was produced by a professional translator (6).

[厚生省が昨年公表した[1]] [人口の将来推計では、[2]] [将来、一・四九九人を最低に、[3]] [その後は上昇に転ずると[4]] [推計していたが、[5]] [早くも予想がはずれる[6]] [見通しとなった。[7]] (4)

[The Ministry of Health and Welfare last year revealed[1]] [population of future estimate according to[2]] [in future 1.499 persons as the lowest[3]] [that after *SAB* rising to turn that[4]] [*they* estimated but[5]] [already the estimate misses a point[6]] [prediction became.[7]] (5)

[In its future population estimates[1]] [made public last year.[2]] [the Ministry of Health and Welfare predicted that the SAB would drop to a new low of 1.499 in the future.[3]] [but would make a comeback after that.[4]] [increasing once again.[5]] [However, it looks as if that prediction will be quickly shattered.[6]] (6)

The labeled spans of text represent elementary discourse units (*edus*), i.e., minimal text spans that have an unambiguous discourse function (Mann and Thompson, 1988). If we analyze the text fragments closely, we will notice that in translating sentence (4), a professional translator chose to realize the information in Japanese unit 2 first (unit 2 in text (4) corresponds roughly to unit 1 in text (6)); to realize then some of the information in Japanese unit 1 (part of unit 1 in text (4) corresponds to unit 2 in text (6)); to fuse then information given in units 1, 3, and 5 in text (4) and realize it in English as unit 3; and so on. Also, the translator chose to re-package the information in the original Japanese sentence into two English sentences.

At the elementary unit level, the correspondence between Japanese sentence (4) and its English translation (6) can be represented as in (7), where $j \subset e$ denotes the fact that the semantic content of unit j is realized fully in unit e; $j \supset e$ denotes the fact that the semantic content of unit e is realized fully in unit j; $j = e$ denotes the fact that units j and e are semantically equivalent; and $j \cong e$ denotes the fact that there is a semantic overlap between units j and e, but neither proper inclusion nor proper equivalence.

$$
\begin{aligned}
&j_1 \supset e_2; j_1 \cong e_3;\\
&j_2 = e_1;\\
&j_3 \subset e_3;\\
&j_4 \cong e_4; j_4 \cong e_5;\\
&j_5 \cong e_3;\\
&j_6 \subset e_6;\\
&j_7 \subset e_6
\end{aligned}
\tag{7}
$$

Hence, the mappings in (7) provide an explicit representation of the way information is re-ordered and re-packaged when translated from Japanese into English. However, when translating text, it is not only that information is re-packaged and re-ordered; it is also that the rhetorical rendering changes. What is realized in Japanese using an ELABORATION relation can be realized in English using, for example, a CONTRAST or a CONCESSION relation.

Figure 2 presents in the style of Mann and Thomp-

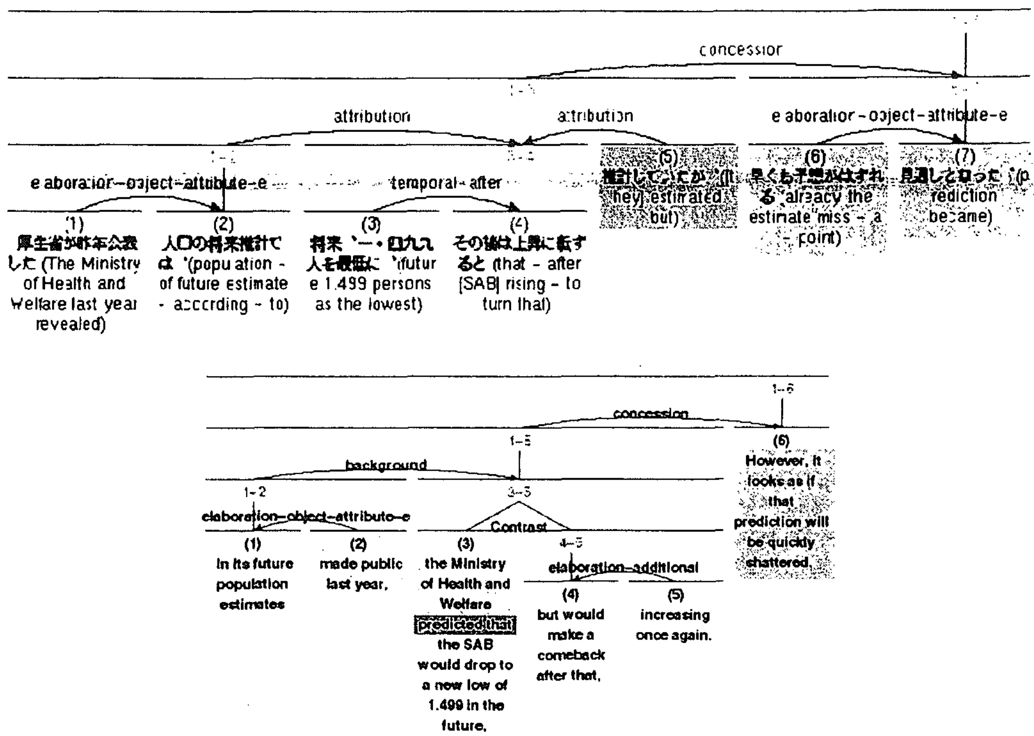

Figure 2: The discourse structures of texts (4) and (6).

son (1988) the discourse structures of text fragments (4) and (6). Each discourse structure is a tree whose leaves correspond to the *edus* and whose internal nodes correspond to contiguous text spans. Each node is characterized by a *status* (NUCLEUS or SATELLITE) and a *rhetorical relation*, which is a relation that holds between two non-overlapping text spans. (There are a few exceptions to this rule: some relations, such as the CONTRAST relation that holds between unit [3] and span [4,5] in the structure of the English text are multinuclear.) The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. Rhetorical relations that end in the suffix "-e" denote relations that correspond to embedded syntactic constituents. For example, the ELABORATION-OBJECT-ATTRIBUTE-E relation that holds between units 2 and 1 in the English discourse structure corresponds to a restrictive relative. We chose to label these relations because we have noticed that they often dominate complex discourse trees, whose elementary units are fully fleshed clauses.

If one knows the mappings at the *edu* level, one can determine the mappings at the span (discourse constituent) level as well. For example, using the elementary mappings in (7), one can determine that Japanese span [1,2] corresponds to English span [1,2], Japanese unit [4] to English span [4,5], Japanese span [6,7] to English unit [6], Japanese span [1,5] to English span [1,5], and so on. As Figure 2 shows, the CONCESSION relation that holds between spans [1,5] and [6,7] in the Japanese tree corresponds to a similar relation that holds between span [1,5] and unit [6] in the English tree (modulo the fact that, in Japanese, the relation holds between sentence fragments, while in English it holds between full sentences). However, the TEMPORAL-AFTER relation that holds between units [3] and [4] in the Japanese tree is realized as a CONTRAST relation between unit [3] and span [4,5] in the English tree. And because Japanese units [6] and [7] are fused into unit [6] in English, the relation ELABORATION-OBJECT-ATTRIBUTE-E is no longer made explicit in the English text.

Assume now that it is the task of an MGEN system to produce from a knowledge base texts (4) and (6). The system will have to select the appropriate information, generate text plans for the two texts, generate sentence plans, and realize them. Should the system generate a text plan having a structure similar to the RST analysis at the top or the bottom of Figure 2? Or something in between? As one can see, the discourse trees in Figure 2 are quite different: they suggest that depending on the output language, text plans should use different relations, different orderings of elementary units, different aggregations across semantic units, etc.

Some researchers may argue that the two RST analyses in Figure 2 are too specific. That they, in fact, correspond to text plans that have been already refined by an aggregation module and arguably, even by a sentence planner. After all, the re-ordering of units 1 and 2 can be explained only in terms of different syntactic contraints in Japanese and English. We agree with such a concern. Nevertheless, as our experiment shows, significant differences across discourse trees are found not only for trees built at the sentence level, but also for trees built at the paragraph and text levels. For these levels, it is difficult to explain the differences in terms of language-specific syntactic constraints. Rather, it seems more adequate to assume that there are significant differences with respect to the way information is organized rhetorically across languages. The experiment described in the next section estimates quantitatively this difference.

3 Experiment

In order to assess how similar discourse structures are across languages, we built manually a corpus of discourse trees for 40 Japanese texts and their corresponding translations. The texts, selected randomly from the ARPA corpus (White and O'Connell, 1994), contained on average about 460 words. We developed a discourse annotation protocol for Japanese and English along the lines followed by Marcu et al. (1999). We used Marcu's discourse annotation tool (1999) in order to manually construct the discourse structure of all Japanese and English texts in the corpus. 10% of the Japanese and English texts were rhetorically labeled by two of us. The agreement was statistically significant ($Kappa = 0.65, \alpha > 0.01$ for Japanese and $Kappa = 0.748, \alpha > 0.01$ for English (Carletta, 1996; Siegel and Castellan, 1988)). The tool and the annotation protocol are available at *http://www.isi.edu/~marcu/software/*. For each pair of Japanese-English discourse structures, we also built manually an alignment file, which specified the correspondence between the *edus* of the Japanese and English texts.

Using labeled recall and precision figures, we computed the similarity between English and Japanese discourse trees with respect to their assignment of *edu* boundaries, hierarchical spans, nuclearity, and rhetorical relations. Because the trees we compared differ from one language to the other in the number of elementary units, the order of these units, and the way the units are grouped recursively into discourse spans, we computed two types of recall and precision figures.

In computing *Position-Dependent* (P-D) recall and precision figures, a Japanese span was considered to match an English span when the Japanese span contained all the Japanese *edus* that corresponded to the *edus* in the English span, and when the Japanese and English spans appeared in the same position linearly. For example, the English tree in Figure 2 is characterized by 10 subsentential spans, which span across positions [1,1], [2,2], [3,3], [4,4], [5,5], [6,6], [1,2], [4,5], [3,5], and [1,5]. (Span [1,6] subsumes 2 sentences, so it is not sub-sentential.) The Japanese discourse tree has only 4 spans that could be matched in the same positions with English spans, namely spans [1,2], [4,4], [5,5], and [1,5]. Hence the similarity between the Japanese tree and the English tree with respect to their discourse structure below the sentence level has a recall of 4/10 and a precision of 4/11 (in Figure 2, there are 11 sub-sentential Japanese spans).

In computing *Position-Independent* (P-I) recall and precision figures, even when a Japanese span "floated" during the translation to a position in the English tree that was different from the position in the initial tree, the P-I recall and precision figures are affected less than when computing Position-Dependent figures. The position-independent figures reflect the intuition that if two trees t_1 and t_2 both have a subtree t, t_1 and t_2 are more similar than if they were if they didn't share any subtree. For instance, for the spans at the sub-sentential level in the trees in Figure 2 the position-independent recall is 6/10 and the position-independent precision is 6/11 because in addition to spans [1,2], [4,4], [5,5], and [1,5], one can also match Japanese span [1,1] to English span [2,2] and Japanese span [2,2] to Japanese span [1,1]. The Position-Independent figures offer a more optimistic metric for comparing discourse trees. They span a wider range of values than the Position-Dependent figures, which enables a finer grained comparison, which in turn enables a better characterization of the differences between Japanese and English discourse structures.

In order to provide a better estimate of how close two discourse trees were, we computed Position-Dependent and -Independent recall and precision figures for the sentential level (where units are given by *edus* and spans are given by sets of *edus* or single sentences); paragraph level (where units are given by sentences and spans are given by sets of sentences or single paragraphs); and text level (where units are given by paragraphs and spans are given by sets of paragraphs). These figures offer a detailed picture of how discourse structures and relations are mapped from one language to the other. Some of the differences at the sentence level can be explained by differences between the syntactic structures of Japanese

Level	Units		Spans		Nuclei		Relations	
	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P
Sentence	29.1	25.0	27.2	22.7	21.3	17.7	14.9	12.4
Paragraph	53.9	53.4	46.8	47.3	38.6	39.0	31.9	32.3
Text	41.3	42.6	31.5	32.6	28.8	29.9	26.1	27.1
Weighted Average	36.0	32.5	31.8	28.4	26.0	23.1	20.1	17.9
All	8.2	7.4	5.9	5.3	4.4	3.9	3.3	3.0
	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P	P-I R	P-I P
Sentence	71.0	61.0	56.0	46.6	44.3	36.9	30.5	25.4
Paragraph	62.1	61.6	53.2	53.8	43.3	43.8	35.1	35.5
Text	74.1	76.5	54.4	56.5	48.5	50.4	41.1	42.7
Weighted Average	69.6	63.0	55.2	49.2	44.8	39.9	33.1	29.5
All	74.5	66.8	50.6	45.8	39.4	35.7	26.8	24.3

Table 1: Similarity of the Japanese and English discourse structures

and English. The differences at the paragraph and text levels have a purely rhetorical explanation.

As expected, when one computes the recall and precision figures with respect to the nuclearity and relation assignments, one also factors in the nuclearity status and the rhetorical relation that is associated with each span.

Table 1 summarizes the results (P-D and P-I (R)ecall and (P)recision figures) for each level (Sentence, Paragraph, and Text). It presents Recall and Precision figures with respect to span assignment, nuclearity status, and rhetorical relation labeling of discourse spans. The numbers in the "Weighted Average" line report averages of the Sentence-, Paragraph-, and Text-specific figures, weighted according to the number of units at each level. The numbers in the "All" line reflect recall and precision figures computed across the entire trees, with no attention paid to sentence and paragraph boundaries.

Given the significantly different syntactic structures of Japanese and English, we were not surprised by the low recall and precision results that reflect the similarity between discourse trees built below the sentence level. However, as Table 1 shows, there are astonishing differences between discourse trees at the paragraph and text levels as well. For example, the Position-Independent figures show that only about 62% of the sentences and only about 53% of the hierarchical spans built across sentences could be matched between the two corpora. When one looks at the nuclearity status and rhetorical relations associated with the spans built across sentences, the P-I recall and precision figures drop to about 43% and 35% respectively.

The differences in recall and precision are explained both by differences in the way information is packaged into paragraphs in the two languages and the way it is structured rhetorically both within and above the paragraph level.

4 How should a multilingual text planner work?

The results in Section 3 strongly suggest that if one is to build text plans in the context of a Japanese-English multilingual generation system, a language-independent text planning module whose output is mapped straightforwardly into sentence plans (Iordanskaja et al., 1992; Goldberg et al., 1994) will not do. The differences between the rhetorical structures of Japanese and English texts are simply too big to support the derivation of a unique text plan, which would subsume both the Japanese- and English-specific realizations. If we are to build MGEN systems capable of generating rich texts in languages as distant as English and Japanese, we would need to use more sophisticated techniques. In the rest of this section, we discuss a set of possible approaches, which are consistent with work that has been carried out to date in the NLG field.

4.1 Use text plan representations that are more abstract than discourse trees

Delin et al. (1994) have shown that although the rhetorical renderings in Figure 1 are non-isomorphic, they are all subsumed by one common, more abstract text-plan representation language that formalizes the procedural relations of *Generation* and *Enablement* (Goldman, 1970). One can conceive of text plans being represented as sequences of actions or hierarchies of actions and goals over which one can identify *Generation* and *Enablement* relations that hold between them. In such a framework, text planning is carried out in a language-independent manner, which is then followed by a rhetorical "fleshing out". (Delin et al. (1994) have shown how Generation and Enablement relations are realized rhetorically in various languages using relations such as PURPOSE, SEQUENCE, CONDITION, and MEANS.)

Bateman and Rondhuis (1997) suggest that the variability present in Delin et al.'s Rhetorical Structure analyses in Figure 1 can be explained by the inadequate mixture of intentional and semantic relations, at different levels of granularity. They propose that discourse phenomena should be accounted for at a more abstract level than RST relations and they present a classification system in terms of "stratification", "metafunction", and "paradigmatic/syntagmatic axiality" that enables one to represent discourse structures at multiple levels of abstraction.

Adopting such an approach could be an extremely rewarding enterprise. Unfortunately, the research of Delin et al. (1994) and Bateman and Rondhuis (1997) cannot be applied yet to unrestricted domains. Generation and Enablement are only two of the abstract relations that can hold between actions and goals. And some texts, such as descriptions, are difficult to characterize only in terms of actions and goals. Building a "complete" taxonomy of such abstract relations and identifying adequate mappings between there relations and rhetorical relations are still open problems.

4.2 Derive a language-independent discourse structure, and then linearize it

Rösner and Stede (1992) and Stede (1999) assume that a discourse representation à la Mann and Thompson imposes no contraints on the linear order of the leaves. For the purpose of multilingual text planning, one can, hence, assume that a language-independent text planner derives first a language-independent rhetorical structure and then *linearizes* it, i.e., transforms it to make it language specific. The transformations that Rösner and Stede have applied concern primarily re-orderings of the children of some nodes and re-assignment of rhetorical relation labels. But given, for example, the significant differences between the discourse structures in Figure 2, it is difficult to envision what the language-independent text plan might look like. It is definitely possible to conceive of such a text plan representation. However, the linearization module will need then to be much more sophisticated: it will need to be able to rewrite full structures, re-order constituents, aggregate across possibly non-adjacent units, etc.

4.3 Implement a text planning algorithm for one language only. For all other languages, devise discourse-tree rewriting modules

In this approach, the system developer assigns a preferrential status to one of the languages that are to be handled by the MGEN system. Let's call this language P. The system developer implements text planning algorithms only for this language. For any other language O, the developer implements a discourse-tree rewriting module capable of rewriting P-specific discourse structures into O-specific discourse structures. When generating texts in language P, the MGEN system works as a monolingual generator. When generating texts in language O, the MGEN system generates a text plan in language P, maps it into language O, and then proceeds further with the sentence planning and realization stages. Marcu et al. (2000) present and evaluate a discourse-tree rewriting algorithm that exploits machine learning methods in order to map Japanese discourse trees into discourse trees that resemble English-specific renderings.

The advantage of such an approach is that the tree-rewriting modules can be also used in the context of machine translation systems in order to re-package and re-organize the input text rhetorically, to reflect constraints specific to the target language. The disadvantage is that, from an NLG perspective, there is no guarantee that such a system could produce better results than a system that implements language-dependent text planning modules.

4.4 Derive language-dependent text plans

Another viable approach is to acknowledge that text plans vary significantly across languages and, therefore, should be derived by language-dependent planners. To this end, one could use both top-down (Hovy, 1993; Moore and Paris, 1993) and bottom-up (Marcu, 1997; Mellish et al., 1998) text planning algorithms. The advantage of this approach is that it has the potential of producing trees that reflect the peculiarities specific to any language. The disadvantage is that only the text planning algorithms are general: the plan operators and the rhetorical relations they operate with are language-dependent, and hence, more expensive to develop and maintain.

4.5 Discussion

Depending on the languages and text genres it operates with, an MGEN system may get away with a language-independent text planner. However, for sophisticated genres and distant languages, implementing a language-independent planner that is straightforwardly mapped into sentence plans does not appear to be a felicitous solution. We enumerated four possible alternatives for addressing the text planning problem in an MGEN system. Each of the approaches has its own pluses and minuses. Which will eventually win in large-scale deployable MGEN systems remains an open question.

References

John A. Bateman and Klaas Jan Rondhuis. 1997. Coherence relations: Towards a general specification. *Discourse Processes*, 24:3–49.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June.

Judy L. Delin, Anthony Hartley, Cécile L. Paris, Donia R. Scott, and Keith Vander Linden. 1994. Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 61–70, Kennebunkport, Maine, June.

J. Delin, D. Scott, and A. Hartley. 1996. Pragmatic congruence through language-specific mappings from semantics to syntax. Technical report, ITRI Research Report ITRI-96-12, University of Brighton.

E. Goldberg, N. Driedger, and R. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

A.I. Goldman. 1970. *A Theory of Human Action*. Prentice Hall, Englewood Cliffs, NJ.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1–2):341–386, October.

L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, and A. Polguere. 1992. Generation of extended bilingual statistical reports. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, pages 1019–1023, Nantes, France.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pages 629–635, Providence, Rhode Island, July 28-31.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, University of Maryland, June 22.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics NAACL-2000*, Seattle, Washington, April 29 – May 3.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 98–107, Niagara-on-the-Lake, Canada, August 5-7.

Johanna D. Moore and Cécile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694.

C. Paris, K. Vander Linden, M. Fischer, A. Hartley, L. Pemberton, R. Power, and D. Scott. 1995. A support tool for writing multilingual instructions. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1398–1404, Montreal, Canada.

Richar Power and Donia Scott. 1998. Multilingual authoring using feedback texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, Montreal, Canada, August.

Ehud Reiter and Chris Mellish. 1993. Optimizing the costs and benefits of natural language generation. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1164–1169.

Dietmar Rösner and Manfred Stede. 1992. Customizing RST for the automatic production of technical manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation; 6th International Workshop on Natural Language Generation*, number 587 in Lecture Notes in Artificial Intelligence, pages 199–214, Trento, Italy, April. Springer-Verlag.

Sidney Siegel and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, second edition.

Manfred Stede. 1999. Rhetorical structure and thematic structure in text generation. In *Working Notes of the Workshop on Levels of Representation in Discourse*, pages 117–123, Edinburgh, Scotland, July 7-9.

J. White and T. O'Connell. 1994. Evaluation in the ARPA machine-translation program: 1993 methodology. In *Proceedings of the ARPA Human language Technology Workshop*, pages 135–140, Washington, D.C. See also *http://ursula.georgetown.edu/*.