Andrej Bekeš,
University of Tsukuba & Jožef Stefan Institute

CDU 801.52 : 809.46-561

RELATEDNESS OF CONTENT AND SENTENCE FORMATION IN JAPANESE[1]

0. INTRODUCTION

Leech (1983: 63–70) distinguishes two kinds of pragmatics, interpersonal pragmatics and textual pragmatics. Our article is concerned with textual pragmatics, specifically with the textual motivations behind a format such as a sentence in Japanese.

Studying spontaneous spoken discourse, Chafe (1980) proposed two units of spoken discourse on the basis of phonetical and intonational criteria, i.e. the "idea unit" and the "intonation sentence". He finds justification for both units in cognitive processes as follows. Idea units, most often verbalized as clauses, are the linguistic expression of cognitive units that Chafe calls "foci of consciousness". A focus of consciousness is a chunk of information small enough to be processed and verbalized in one step. Next, an intonation sentence, consisting usually of several idea units (or sometimes just one) is the verbal expression of a larger cognitive unit, the "center of interest", a chunk of information too large to be verbalized in one step. Concerning the center of interest, Chafe puts forward the following hypothesis.

> Spontaneous spoken language then suggests the existence of some sort of cognitive entity which I am calling a center of interest and which corresponds roughly to what is expressed in a linguistic sentence.
>
> (Chafe 1980: 29)

Applying the above hypothesis to written language, I tried to measure content relatedness between clauses (Bekeš 1985, 1987). It follows from Chafe's hypothesis that clauses coinciding inside an intonational sentence belong to the same hypothetical cognitive unit, the center of interest, and should therefore be more closely connected in regard to their cognitive content than clauses not appearing in different intonation sentences. Because of the close connection that is supposed to exist between centers of interest and linguistic sentences, the above consequence could be expected to hold for sentences of the written language as well. On the basis of such reasoning I introduced an empirical measure of content relatedness (*yuuensei*). I defined content relatedness on the basis of paraphrases of texts, written in what Givón (1979) calls pragmatic mode, where each sentence corresponds roughly to a clause. The frequency of

[1] This paper is based on an earlier version published in Japanese as Bekeš (1992).

any two clauses from the original coinciding within the same sentence in paraphrases was taken as the quantitative measure of their content relatedness.

Chafe (1987), while admitting difficulties encountered when we try to define the concept of sentence precisely, offers a revision of his first hypothesis, supporting it by an analysis of new spoken discourse material. The new hypothesis considers a sentence to belong more to the realm of rhetorics than to have a clear-cut cognitive basis.

> The function of sentences in spoken languages is intriguing and problematic... There is a useful distinction to be made between those linguistic units which are determined by basic cognitive phenomena such as memory and consciousness and those which result from passing decisions regarding coherence and rhetorical effect. In the former, cognitively determined category, belong intonation units, extended clauses and paragraphs. Sentences on the other hand seem to belong to the category of phenomena which are under more rhetorical control, and are more independent of cognitive constraints.
>
> (Chafe 1987: 46)

In his new hypothesis he relativizes and complements his first hypothesis. That is, the cognitive basis of a sentence is a matter of degree while the rhetorical considerations may actually be more important.

Chafe bases his hypotheses on a qualitative analysis of spoken discourse. In this paper I attempt a quantitative verification of Chafe's propositions. I analyze written paraphrases elicited from Japanese native speakers. In section 1 I describe the experiment, centered around written paraphrases of a short text written in pragmatic mode. In section 2 I examine content relatedness between clauses of the input text as reflected in intuitions of people who participated in the paraphrase experiment. In section 3 I analyze the coincidence of clauses within the same sentence in paraphrases. In section 4 I examine the connection between content relatedness and the coincidence of clauses within the same sentence. In the last section I discuss the results.

1. THE EXPERIMENT

In order to find out how clauses are combined into sentences I use the same paraphrase method as in Bekeš (1985, 1987), except that the method and emphasis of analysis, as well as the lines of interpretation are to some extent different. The paraphrase experiment was done with 45 first year students of the University of Tsukuba College of Physical Education.

Clauses seem to be written language counterparts of Chafe's idea units (Chafe 1980). Therefore the input text used in paraphrases is realized under the restriction that one sentence should consist of one clause. To observe how clauses were combined into sentences, participants in the experiment had to complete three tasks.

The first task, task A (below PARAPHRASE EXPERIMENT) was to paraphrase the input text as a news article.[2]

Next, in order to control the participants' reading of the input text, they had to write a short summary of the input text in task B.

Finally, in order to find out the participants' intuitive judgement of content relatedness between input text clauses, in task C (below MARKING EXPERIMENT), participants had to mark those clauses in the input text they thought to be related in content within the context of the whole input text.

The input text used in the experiment is the text shown below together with its English translation.

"Kaisyain no zisatu"

1. Sakuzitu no yugata no koto desita.
2. Kanagawa-ken XX-si no XX-yama no tyuuhuku de otoko no hito ga kubi wo tutte imasita.
3. Sono otoko wa sinde imasita.
4. Aru hito ga imohori ni dekakemasita.
5. Sono hito ga sono sitai o mitukemasita.
6. Sono koto o XX-keisatusyo ni todokemasita.
7. Keisatu wa kore o sirabemasita.
8. Sitai wa Aiti-ken no XX-si no kaisyain O.san desita.
9. Sore wa sebiro no nemu kara wakarimasita.
10. Kyonen no kugatu, tyoonan ga XX-sinai de ziko o okosimasita.
11. Tyoonan ga aru onna no hito ni nikagetu no zyuusyuo o owasemasita.
12. Zyosei wa kyonen no kure ni siboo simasita.
13. O. san wa sore o sirimasita.
14. Kare wa sono mama kaisya o sootai simasita.
15. Ie ni mo kaerimasendesita.
16. Yukuehumei ni narimasita.
17. Kazoku ga soo hanasite imasita.
18. Sikasi sono onna no hito no siboo gen'in wa sinhuzen desita.
19. Ziko to wa tyokusetu kankei arimasendesita.
20. Sore nanoni, O.san wa koo omoimasita.
21. Ziko ga gen'in da, to.
22. Onna no hito ga sorede siboo sita, to.
23. O.san wa sekinin wo kanzimasita.
24. Zisatu simasita.
25. Keisatu de wa izyoo no yoo ni mite imasu.

"Employee's suicide"

1. It happened last night
2. In a forest in Kanagawa prefecture there was a man hanging.
3. The man was dead.
4. A peasant went to dig edible roots.
5. This peasant found the corpse.

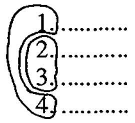
[2] Corresponds to paraphrase experiment (simple sentences to complex sentences) in Bekeš (1985/87: Ch. 4).

6. He reported this to the prefectural police.
7. The police investigated the case.
8. The corpse was of an employee, Mr. O. from XX city in Aichi prefecture.
9. This was found out from the name on the jacket.
10. In September last year Mr. O.'s eldest son caused a traffic accident in his hometown.
11. The eldest son inflicted heavy injuries upon some woman.
12. This woman died last fall.
13. Mr. O. learned about this.
14. He immediately left his office.
15. He did not even go home.
16. He became missing.
17. The family told this.
18. Actually, the cause of the woman's death was a heart trouble.
19. It was not directly connected with the accident.
20. On the other hand Mr. O. thought like this.
21. That the accident was the cause.
22. That the woman died because of this.
23. Mr. O. felt responsible.
24. He killed himself.
25. The above is the police view.

INSTRUCTION FOR THE TASK A:
Without taking away or adding to the content, paraphrase the above text as an objective news report such as you find in newspapers

INSTRUCTION FOR THE TASK B:
Write a short summary of the above text.

INSTRUCTION FOR THE TASK C:
Among the 25 sentences constituting the above text, mark those that you consider to be related in their content within the overall context of the whole text following the example below.



2. INPUT TEXT CLAUSES AND THE MARKING EXPERIMENT

As a result of the marking experiment, any pair of input clauses was assigned value 1 for each participant who marked the pair as related in content. Total score for each pair was represented in a matrix, where rows and columns represent input clauses and each element of the matrix represents the total marking score of the corresponding input clause pair, or, in other words, the frequency of any such pair being marked as related. The matrix is shown below. Since it is symmetric, only the lower half is shown.[3]

[3] In Bekeš (1987) this matrix was interpreted as a kind of numerical measure of content relatedness (*yuuensei*

→ *input clause No.*

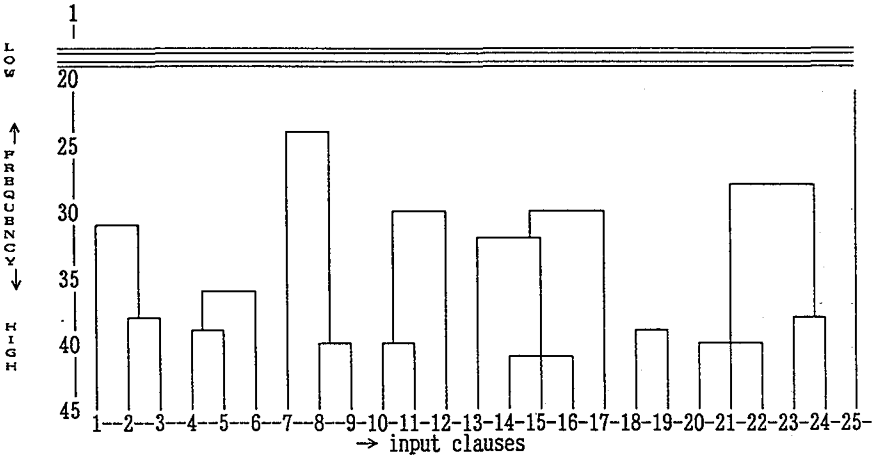
↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	31																							
3	30	38																						
4	5	6	6																					
5	4	5	5	39																				
6	4	5	5	35	36																			
7	1	1	1	2	1	6																		
8	2	2	1	1	1	2	24																	
9	2	2	1	1	1	22	3	40																
10	0	0	0	0	0	0	2	3	3															
11	0	0	0	0	0	0	1	1	1	40														
12	0	0	0	0	0	0	1	1	1	32	30													
13	0	0	0	0	0	0	1	1	1	17	15	21												
14	0	0	0	0	0	0	1	1	1	9	9	12	32											
15	0	0	0	0	0	0	1	1	1	9	9	14	31	41										
16	0	0	0	0	0	0	1	1	1	9	9	14	31	41	41									
17	0	0	0	0	0	0	1	1	1	9	9	11	21	29	30	27								
18	0	0	0	0	0	0	1	1	1	4	4	8	6	7	7	7	9							
19	0	0	0	0	0	0	1	1	1	4	4	7	6	7	7	7	7	39						
20	0	1	1	0	0	0	1	1	1	4	4	3	4	4	4	4	4	17	16					
21	0	1	1	0	0	0	1	1	1	4	4	3	4	4	4	4	4	16	16	40				
22	0	1	1	0	0	0	1	1	1	4	4	3	4	4	4	4	4	15	15	40	40			
23	0	1	1	0	0	0	1	1	1	4	4	3	5	4	4	4	4	11	11	26	28	27		
24	0	1	1	0	0	0	1	1	1	3	3	2	4	3	3	3	3	9	9	23	24	24	38	
25	0	0	0	0	0	0	1	1	1	3	3	2	4	3	3	3	3	4	4	5	5	5	9	10

The matrix itself already reveals an internal structure. High frequency scores are centered in blocks along the diagonal, implying grouping of clauses into chunks with relatively strong content relatedness within such chunks. However, we need a subtler method to arrive at more valid conclusions. Since the matrix represents a kind of similarity matrix, with the marking frequencies standing for similarity measure, it is possible to apply one of the cluster analysis methods and see if any clauses tend to cluster together. Applying the minimal distance linkage method (cf. Anderberg 1973) the following clustering diagram was obtained.[4]

I /relatedness I/).

[4] Levelt (1974: Ch. 2) used the same method to determine semantic relatedness between immediate constituents of a sentence.

Graph 1: Clustering diagram of the marking experiment



The vertical axis of the diagram represents the observed frequency of clause pairs being marked as related. It shows at which level input clauses, listed along the horizontal axis, were merged together in clusters. Terminal or near terminal clusters merge at a higher frequency and are therefore more important than higher level clusters. Therefore we shall limit ourselves to the terminal clusters, containing only clauses (marked by "()"), to the level 2 clusters, containing clauses as well as terminal clusters (marked by "[]"), and to the level 3 clusters, containing besides clauses also terminal clusters as well as clusters of the level 2 (marked by "{ }"). We shall also limit ourselves to clusters, merged at the level higher than 20 (this frequency represents the approximate median for the whole matrix). The input clauses then appear merged in clusters as follows:

(2) [1 (2 3)] [(4 5) 6] [7 (8 9)] [(10 11) 12] {[13 (14 15 16)] 17} (18 19) [(20 21 22)(23 24)] 25

The clauses contained in these clusters are those marked by the participants as being most related.

What then is the intuition which we may suspect behind such grouping? Using the notion of TOPIC CONTINUITY as proposed by Givón (1983, 1989) (intuitively it corresponds to the entity being talked about), we group clauses sharing the same topic entity in the same cluster. Thus we arrive at the following grouping of the input clauses.

(3)	(1)	(2 3)	(4 5 6)	(7)	(8 9)	(10 11)	(12)	(13 14 15 16)	(17)	(18 19)	(20)	(21 22)	(23 24)	(25)
	T	O	P	K	O	S	V	O	F	V	O	V	O	K

Topic entities:
T = time; O = Mr. O.; P = person who discovered O.; K = police; S = O.'s son; V = victim; F = O.'s family

Now let us examine the content of the clusters in (3). Cluster (1) is a single clause, specifying the time when (2 3) has happened. Cluster (2 3) describes O.'s suicide. Cluster (4 5 6) describes how O.'s corpse has been discovered, including the reason why the person who discovered O.'s corpse went there and his subsequent action. Single clause cluster (7) provides information about police action. Cluster (8 9) tells what the results of the investigation were, providing identity of the corpse. Cluster (10 11) tells about the accident caused by O.'s son and its consequences. Cluster (12) includes a single clause, introducing the victim of the accident. Since the victim does not appear again in near vicinity, this clause stands out isolated. Clauses (13 14 15 16) tell us about O.'s actions after he heard about their death of the victim, until he became missing. Single clause cluster (17) provides us with the source of information of the previous cluster, i.e. O.'s family. Cluster (18 19) tells the reason of the victim's death. Single clause cluster (20) specifies the source and modality of information appearing in the subsequent cluster (21 22), i.e. about the supposed reason of the victim's death. Clauses (23 24) include the information about O.'s actions leading to committing a suicide. And the last cluster, again a single clause, provides the source and modality of the information, specified in clauses from 18 to 24.

In (3) all the single clauses, except clause 12, are clauses related to evidentiality (cf. Chafe 1986), i.e. clauses specifying the source and type, i.e. fact, supposition etc, of information (i.e. clauses 7, 17, 20, 25) or specifying background information (i.e. clause 1).

It is interesting to note that clauses 1, 7, 17, and 25 appear also in the marking experiment, (GRAPH 1) as relatively loosely attached to more strongly merged clusters of other clauses. In other words, these single clauses from (3) appear in GRAPH 1 as merged with other clusters at the level 2 or 3 or even higher, meaning a relatively looser association. Common point of all these clauses (i.e. 1, 7, 17, 25) is that they are simultaneously related in content to several other clauses at the same time, though none of the relations involves topic continuity.[5]

There remain the cases of behaviour of clauses 12 and 20 in the marking experiment. Clause 12 is actually not the only clause including the potential topic entity "victim". "Victim" appears also in clause 11. The reason why it does not appear as topic entity is that there is a more powerful candidate for the same role in clauses 10 and 11. It is "O.'s son", appearing as subject (definition of subject is according to Sibatani 1978) in clause 10 and as deleted (or better, nonexpressed) subject in clause

[5] A possible reason why in the marking experiment such clauses were left single more often than other clauses seems to lie in the fact that each participant had at his/her disposal only one level to mark content relatedness of input clauses. Where there was a possibility of choice, strong content relations, based on topic content, seem to have had more chance to be marked explicitly than the weaker relations, described above.
A propos the tendency of such clauses to be left single, in the paraphrase experiment we also observe a similar phenomenon. The reason here is that the one-dimensional chain of linguistic signs does not provide an easy means for specifying such hierarchical relations. For treatment of these phenomena in Japanese see Bekeš (1985/87, ch. 5).

11. Deletion is a very powerful cohesive and topic continuity marker and therefore "O.'s son" is preferentially interpreted as the topic entity in this case.

The reason for clause 20, which is also providing evidentiality information, to appear in (2) merged closely with its content clauses 21 and 22, is perhaps that the whole segment of the text is embedded as the content of police reasoning within the range of clause 25, and the compactness of the cluster containing clauses 20, 21, 23 is thus an expression of relatively higher content relatedness of these clauses in regard to clause 25.

As we have seen, each marking experiment participant had at his/her disposal only one level to mark the content relatedness of input clauses. From the similarity between clusters in (3) and (2) we may assume that participants' judgement had to be to a great extent based on the topic continuity that they intuitively observed in the input text. Thus we may understand the hierarchical structure of content seen in clause clusters resulting from the marking experiment to be at least partially a consequence of topic continuity in the input text.

From GRAPH 1 we can also see, that the input text is segmented into several larger chunks. The deepest discontinuity appears between clusters (1–9), (10–17), (18–19) and (20–25). Again, within the cluster (1–9) there is a rather deep discontinuity between clusters (1–6) and (7–9). Cluster (1–9) as a whole is merged at the frequency level 6, low compared to the maximal frequency of 45.

3. THE PARAPHRASE EXPERIMENT

Next we shall consider the results of the paraphrase experiment (task A). The purpose of this experiment was to observe how participants will merge input clauses into complex sentences in their paraphrases. Again, each occurrence of a pair of input clauses in the same sentence in one of the paraphrases was counted as 1 in the count of total coincidence score for each possible pair of input clauses.[6] As in the case of the marking experiment, the total scores were assembled in a symmetrical coincidence matrix, shown here in the upper diagonal form.

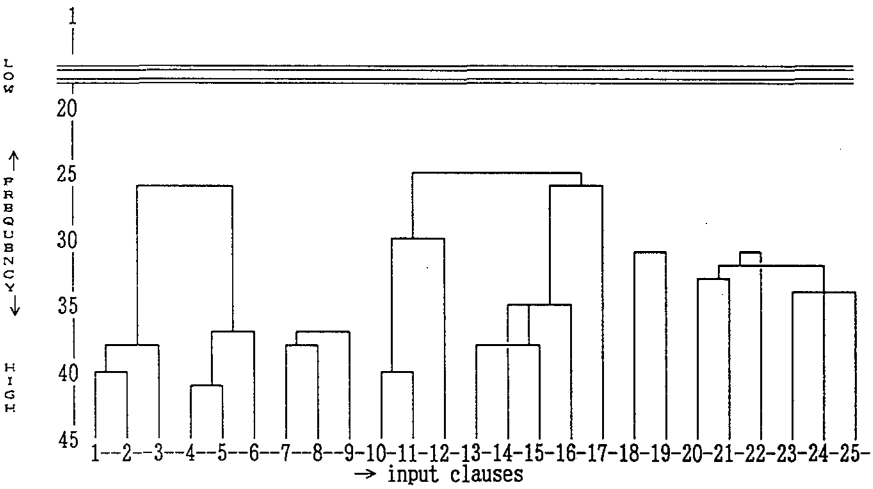
[6] I discussed the issue of identification of the input clauses in paraphrases at length in Bekeš (1985/87). The accuracy test, based on independent identification performed by two different persons gives about 97 % accuracy for the experiment described here. This is well within the limits of statistical fluctuations.

→ *input clause No.*

↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	40																							
3	38	38																						
4	20	21	22																					
5	25	26	24	41																				
6	18	19	19	37	37																			
7	0	2	1	2	2	5																		
8	0	2	1	3	3	6	38																	
9	0	0	1	2	2	5	35	37																
10	1	1	0	1	0	0	5	1	1															
11	1	1	0	1	0	0	5	1	1	40														
12	1	1	0	0	0	0	4	1	1	30	30													
13	1	1	0	0	0	0	2	0	0	17	17	23												
14	1	1	0	0	0	0	1	0	0	14	14	25	33											
15	1	1	0	0	0	0	1	0	0	14	14	19	38	34										
16	1	0	0	0	0	0	1	0	0	16	16	20	33	35	34									
17	0	0	0	0	0	0	1	0	0	13	13	15	23	24	25	26								
18	0	1	0	0	0	0	1	0	0	5	4	8	4	6	5	6	6							
19	0	0	0	0	0	0	1	0	0	2	2	2	1	2	1	2	5	31						
20	1	1	0	0	0	0	1	0	0	5	5	7	4	6	6	6	4	17	15					
21	1	1	0	0	0	0	1	0	0	5	5	7	4	6	6	6	4	16	14	33				
22	1	0	0	0	0	0	0	0	0	2	2	2	2	2	2	1	3	12	13	28	30			
23	1	0	0	0	0	0	0	0	0	3	3	5	3	2	2	1	3	11	12	28	28	27		
24	1	5	2	1	1	1	2	1	0	7	7	8	5	5	4	4	4	16	15	32	31	28	34	
25	1	0	0	0	0	0	0	0	0	4	4	5	3	3	2	2	3	14	13	28	27	26	31	34

In this matrix we can again observe the internal structure, where the highest coincidence frequencies are centered in blocks along the diagonal, already signaling the presence of several large clusters of clauses. To extract finer clusters of input clauses, which tended to appear within the same sentence in paraphrases, cluster analysis (minimal distance linkage method) was applied to the matrix. The resulting clustering diagram is shown below.

Graph 2: Clustering diagram of the paraphrase experiment



The above clustering diagram even at the first glance resembles that of the marking experiment. There are small discrepancies among terminal clusters or clusters on the levels immediately above terminal. We will discuss these below. But there are no discrepancies such as having clusters with clearly defined boundaries (i.e. merged at a significantly higher frequency than the frequency connecting it to another cluster) in one diagram which would be split among two or more clusters in the other diagram. In other words, all the great groupings mentioned at the end of section 2, except the clause 25, are the same in both diagrams. At the level of terminal clusters, clusters of level 2 and 3, the situation is as follows.

(4) [(1 2) 3] [(4 5) 6] [(7 8) 9] [(10 11) 12] {[(13 15) 14 16] 17} (18 19) {[(20 21) (23 24 25)] 22}

Differences between (2) and (4) appear in the following clusters:

[(1 2) 3], [(7 8) 9], [(13 15), 14 16].

With those at the terminal level, level 2, and level 3:

{[(20 21) (23 24 25)] 22}.

These differences appear great at the first glance, but a closer scrutiny reveals that it is not so. With the exception of the cluster {20–25}, level 2 clusters coincide. Again, within them, the difference of merging frequency of terminal clusters and level 2 clusters appears to be very small: it is 2 in the case of clusters (1 2) and 3, it is 1 in the case of (7 8) and 9, and it is 3 in the case of (13 14), 15 and 16. This means that only 1–3 participants among 35 to 40 included particular clauses in the same sentence while excluding the others. Such small differences, compared with the frequencies of merger of the terminal level and level 2 (i.e. from 35 to 40) do not appear to be statistically significant and may be the result of a statistical variation. This is further supported by the results in Bekeš (1985/87), where terminal and level 2 clusters resulting from the paraphrase experiment coincide with those of the marking experiment. Thus the above result may be seen as a strong merging trend at level 2.

The case of cluster {20-25} is the same, except that the merging frequencies within the whole cluster vary within the range from 31 to 34, again a small difference compared with the total frequency. Because of the larger number of clauses, the whole cluster is merged at level 3.

The trend of level 2 or even level 3 clusters appearing within the narrow range of frequencies is connected with an extended use of syntactic means in the paraphrase experiment. In Bekeš (1985/87: 92) clustering diagrams for the paraphrase experiment of two groups are shown. One group, first year senior high school students of a less prestigious school shows less cluster integration than the other group, second year high school students from another, more prestigious school. The latter group results seem to have a comparable degree of integration to that of the paraphrase experiment used here, with participants being first year university students. These differences seem to be connected with the developing ability to master the writing medium and are an interesting field to explore by themselves.

The reason for an overall similarity between paraphrase and marking experiment results lies most probably in the fact, as was pointed out in Bekeš (1987: 169–172), that the same factor has been underlying participants' activity in both experiments, namely some intuitive perception of content relatedness between input clause pairs on the part of experiment participants.

In Bekeš (1985/87) I argued for the paraphrase experiment on the basis of Chafe (1980) hypothesis about the nature of sentence, and interpreted resulting clusters (such as those in GRAPH 2) in the light of this hypothesis as a hierarchically organized content constituent structure. The observed similarity of clusters in both experiments justifies such an interpretation, as was already pointed out in Bekeš (1987). Further, in the light of our observations in sections 2 and 3, we may say that globally, topic continuity appears to be the factor behind determining the perception of content relatedness as well as the integration of clauses into the same sentence.

Yet the correspondence observed at the cluster level is a global correspondence, pertaining to the whole population of participants. In Bekeš (1987), where I tried to put my initial hypothesis concerning the paraphrase experiment, mentioned in the previous paragraph, on firmer footing by directly investigating content relatedness intuitions, this was the only possible conclusion, since the group involved in the marking experiment was different from the group participating in the paraphrase experiment.

In the present study I am trying to verify the relationship between the two experiments on the level of each individual participant and thereby elucidate the motivation for the sentence format in written language.

4. CONNECTION BETWEEN THE MARKING EXPERIMENT AND THE PARAPHRASE EXPERIMENT RESULTS

There are 300 possible pairs of input clauses in our text but this does not mean that for every such pair the frequency of coincidence within some paraphrased sentence or the frequency of being marked as related carry statistical significance. In order to reduce the unnecessary load of work, we shall here limit our discussion to the most significant level of clusters, i.e. terminal clusters and clusters including clauses and terminal clusters.

In the paraphrase experiment it is reasonable to consider participants' intuition about content relatedness of clauses as an independent variable and actual usage of clauses in paraphrase sentences as dependent variable. Therefore we shall choose clusters obtained from the marking experiment for the departure point. These clusters are as follows.

(5) [1 (2 3)] [(4 5) 6] [7 (8 9)] [(10 11) 12][13 (14 15 16)] [18 19] (20 21 22) (23 24) (25)

Input clauses 17 and 25 are omitted, because they merge with other clauses at the level higher than the first two levels, chosen here.

Table 1 shows the relation between marking content relatedness and coincidence of clauses within the same sentence.

Table 1: Marking of content relatedness and coincidence within the same sentence (terminal and level 2 combined)

		coincidence within the same sentence	
		yes	no
marking	marked	585	100
	unmarked	24	18

χ^2=21.3, p<0.001

The table shows considerable correlation between the two experiments (χ^2 test). To see the dynamics better, we shall compare also correlation within the terminal level clusters (TABLE 2) and between second level clusters (TABLE 3).

Table 2: Marking of content relatedness and coincidence within the same sentence (terminal clusters)

		coincidence within the same sentence	
		yes	no
marking	marked	384	60
	unmarked	3	0

χ^2: computation impossible

TABLE 2 shows that there is proportionally about the same amount of clause pairs in terminal clusters, which are marked as related but which do not appear together in the paraphrases as in TABLE 1. But the proportion of unmarked clauses that appeared together in sentences diminishes greatly compared with TABLE 1. Because of the low frequencies in the "unmarked" line, correlation for this case cannot be computed.

Table 3: Marking of content relatedness and coincidence within the same sentence (level 2 only)

		coincidence within the same sentence	
		yes	no
marking	marked	201	40
	unmarked	21	18

χ^2=16.1, p<0.001

TABLE 3 shows that second level clusters (i.e. those with clauses more loosely marked as related than clauses in terminal clusters) contribute proportionally more to the frequencies in the "unmarked" line as compared to the frequencies in the "marked" line (overall 16.3: 1, terminal clusters 148: 1, second level clusters 6.2: 1). At the same time, the proportion between "coincidence" and "noncoincidence" cases in the "marked" line does not change so drastically (overall 5.85: 1, terminal clusters 6.4: 1 and second level clusters 5: 1). At the same time, TABLE 3 still exhibits strong correlation between marking and coincidence.

From the above three tables it seems that depending on the level of clustering, about 10 %–20 % of marked clause pairs do not end in the same sentence in the paraphrases. This and the coincidence of unmarked pairs, though proportionally much lower, tell us that when integrating clauses within the same sentence, there must be also some other factor at work, besides those purely semantic or cognitive considerations such as topic continuity, factors that, as we have seen, are reflected in marking relatedness.

5. CONCLUSION

In section 3 we saw that the overall trend in grouping clauses within sentences seems to coincide with the trend for the clauses to be marked as related within the context. In section 4 we verified this trend on the level of individual participants. As we have seen in section 2, topic continuity seems to be the prevailing factor, influencing marking of clause relatedness. Chunks of information sharing the same topic seem to be either a series of connected events, sharing the same principal entity or a description of a situation concerning the same principal entity.

Such chunks seem to bear information similar to what Chafe hypothesized as a *center of interest* in spoken discourse. But, as we have seen in section 4, cognitive factors connected with "marking" do not account for the whole phenomenon of sentence formation. Several participants, while still paraphrasing, used in their paraphrases the same 1 clause – 1 sentence strategy. In such cases content relatedness between clauses was signalled not so much by cohesive means that operate exclusively within the sentence (i.e. syntactic means) as by cohesive means operating on a wider scope (anaphora, ellipsis, lexical cohesion etc). The choice of cohesive means for a particular realization of a text indeed seems to depend on stylistic and/or rhetoric considerations, and last but not least, on writer's ability or skill. For example, the choice of syntactical means seems to be connected with more condensed style, or sometimes also with ideological considerations (exemplified by the twisted style of bureaucratic language, cf. Kress & Hodge 1979).

These considerations may account for noncoincidence of marked clause pairs. As I also pointed in Bekeš (1987), overt signalling of content relatedness by syntactical means tends to decrease as the frequency of marked content relatedness decreases. On the other hand, cases where clause pairs were unmarked but coincided within the same sentence in paraphrase texts seem to be connected with the way how syntax operates within a particular language. An attempt to clarify this question using the same experiment material is given in Bekeš (1991).

The above analysis was done on Japanese language material, but it seems to be valid for other languages as well (cf. Bekeš 1992). At the end we may add, that the great role played by intuitions of content relatedness in sentence formation is just another example of iconicity working in syntax. Here semantic proximity of clauses is signalled by their spatial proximity within the sentence while their semantic distance is signalled by the means of formal boundaries delimiting them one from another.

Acknowledgements. I am grateful to all the University of Tsukuba students who participated in this study, and to Jožef Stefan Institute from Ljubljana for letting me on a long sabbatical leave in Japan. Without both the present study would never have materialized. I would also like to thank Kristina Hmeljak for checking my English.

Literature:

Anderberg, M.R. (1973) *Cluster analysis for applications*. Academic Press. N.Y.

Bekeš, A. (1992) Bun no kesei to setu no naiyooteki tunagari (formation of sentence and content relatedness among clauses), in Bunka gengogaku hensyuu iinkai (editorial comitee) *Bunka gengogaku: sono teigen to kensetu* (linguisties and culture: proposale and contributions). Sanseido. Tokyo.

— — (1991) Syuusyoku gozyun to tekusuto no maikuro koozoo – kuroozu kumiawase ni okeru nitiei hikaku (word order in noun modification and local structuring of text – a Japanese English comparative study in clause combining), 4th conference on Japanese language teaching executive committee (ed.) *Dai 4 kai nihnogo renraku kaigi soogoo hookokusyo – tooooken no 'minsyuuka' to nihongo kyooiku* (report from the 4th conference on Japanese language teaching – 'democratization in East Europe and the Japanese language teaching), Lodz, Tokyo.

— — (1989) Kohezivnost v besedilu (cohesion in text). *SOL*. 4–2: 1–16. Zagreb.

— — (1987) *Tekusuto to sintakusu* (text and syntax). Kurosio. Tokyo.

— — (1985) *Tekusuto to sintakusu: nihongo ni okeru kohiizyon no zikkenteki kenkyuu* (text and syntax: experimental study of cohesion in Japanese). PhD dissertation, University of Tsukuba.

Chafe, W.L. (1987) Cognitive constraints on information flow. In R. Tomlin (ed.).

— — (1986) Evidentiality in English conversation and academic writing. In W. Chafe and J. Nichols (eds.).

— — (1980) The deployment of consciousness. In Chafe (ed.).

— — (1979) The flow of thought and the flow of language. In T. Givón (ed.).

— — (1977) Creativity in verbalization and its implications for the nature of stored knowledge. In Freedle, R.O.(ed.).

— — (ed.) (1980) *The pear stories*. Ablex. Norwood. N.J.

Chafe, W. and J. Nichols (eds.) *Evidentiality: the linguistic coding of epistemology*. Ablex. Norwood. N.J.

Dik, S.C. (1978) *Functional grammar*. North-Holland. Amsterdam.

Freedle, R.O.(ed.) (1977) *Advances in discourse processes vol. I: Discourse production and comprehension*. Ablex. Norwood N.J.

Givón, T. (1989) *Mind, code and context*. Lawrence Erlbaum. Hillsdale N.J.

— — (1983) Topic continuity in discourse: an introduction, in T. Givón (ed.).

— — (1979) *On understanding grammar*. Academic Press. N.Y.

— — (ed.)(1983) *Topic continuity in discourse*. Benjamins. Amsterdam.

— — (ed.)(1979) Syntax and semantics vol XII: Discourse and syntax. Academic press. N.Y.

Haiman, John (1985) *Natural syntax*. Cambridge University Press. Cambridge.

Kress, G & R. Hodge (1979) *Language as ideology*. Rutledge & Kegan. London.

Kuno S. (1978) *Danwa no bunpoo* (grammar of discourse). Taisyuukan syoten. Tokyo.

Leech, G. (1983) *Principles of pragmatics*. Longman. London.

Levelt, W.J.M. (1974) *Formal grammars in linguistics and psycholinguistics, vol. I–III*. Mouton. The Hague.

Sibatani M. (1978) *Nihongo no bunseki* (an analysis of Japanese). Taisyuukan syoten. Tokyo.

Tomlin, R. (ed.) (1987) *Coherence and grounding in discourse*. John Benjamins. Amsterdam.

Povzetek

VSEBINSKA POVEZANOST IN OBLIKOVANJE POVEDI V JAPONŠČINI

V članku obravnavam vsebinsko povezanost med stavki (clauses) kot pragmatično motivacijo za oblikovanje sestavljene povedi v japonskem besedilu. V ta namen uporabljam metodo ankete in parafraziranja ob sodelovanju 45 govorcev japonskega jezika.

Prvotno besedilo, realizirano v prostih stavkih so udeleženci parafrazirali v besedilo, realizirano v sestavljenih povedih. Na osnovi ankete, v kateri so udeleženci označili intuitivno dojeto vsebinsko povezanost med stavki v prvotnem besedilu sta bila dobljena matrika vsebinske povezanosti in drevo hierarhije vsebinske povezanosti med posameznimi stavki. Iz primerjave med hierarhijo vsebinske povezanosti s prvotnim besedilom sledi, da je dojemanje vsebinske povezanosti v tesni zvezi s tematskimi verigami. Iz primerjave med zaznamovanjem vsebinske povezanosti med stavki in njihovim pojavljanjem znotraj iste povedi v posamičnih parafrazah sledi, da obstaja vidna korelacija med zaznamovanjem vsebinske povezanosti in pojavljanjem stavkov znotraj ene povedi.

Motivacija za oblikovanje povedi torej sloni na ikoničnosti. Semantična bližina med stavki se kodira v njihovi prostorski bližini med drugim tako, da se pojavijo znotraj iste povedi. Njihova semantična oddaljenost pa se kodira s formalnimi mejami med posamičnimi povedmi.