Magical Number Seven Plus or Minus Two: Syntactic Structure Recognition in Japanese and English Sentences

Masaki Murata, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara

Communications Research Laboratory, MPT,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan,
{murata,uchimoto,qma,isahara}@crl.go.jp,
WWW home page: http://www-karc.crl.go.jp/ips/murata

Abstract. George A. Miller said that human beings have only seven chunks in short-term memory, plus or minus two. We counted the number of bunsetsus (phrases) whose modifiees are undetermined in each step of an analysis of the dependency structure of Japanese sentences, and which therefore must be stored in short-term memory. The number was roughly less than nine, the upper bound of seven plus or minus two. We also obtained similar results with English sentences under the assumption that human beings recognize a series of words, such as a noun phrase (NP), as a unit. This indicates that if we assume that the human cognitive units in Japanese and English are bunsetsu and NP respectively, analysis will support Miller's 7 ± 2 theory.

1 Introduction

George A. Miller suggested in 1956 that human beings have only seven chunks[1] in short-term memory, plus or minus two [6]. We counted the number of *bunsetsus* (phrases) whose modifiees are undetermined in each step of an analysis of the dependency structure of Japanese sentences and which therefore must be stored in short-term memory, using the Kyoto University corpus [3]. (The Kyoto University corpus is a syntactic-tagged corpus collected from editions of the Mainichi newspaper.) The number was roughly less than nine, that is, the upper bound of Miller's 7 ± 2 rule. This result supposes that bunsetsus whose modifiees are not determined are stored in short-term memory. For the Kyoto University corpus, the number of stored items was less than nine. This result supports Miller's theory. We made a similar investigation of English sentences using a method described by Yngve [8]. We assumed that human beings recognize a series of words, such as a noun phrase (NP), as a unit and found that the required capacity of short-term memory is roughly less than nine.

[1] A chunk is a cognitive unit of information.

2 Short-term memory and the 7 ± 2 theory

Miller said that human beings have only seven chunks in short-term memory, plus or minus two, because the results of various experiments on words, tones, tastes, sight organs indicated approximately seven. The "plus or minus two" indicates an individual-based variation[2].

Although the research on the 7 ± 2 theory belongs to the field of psychology, it can be applied to the field of engineering. In sentence generation, for example, a sentence that exceeds the seven plus or minus two capacity of short-term memory is difficult to understand, so sentences are generated that do not exceed this upper limitation [8]. In human-interface systems, only about seven plus or minus two objects are displayed at one time because if more pieces of information are given, humans have trouble recognizing the images. Research on the 7 ± 2 theory is useful not only for the scientific investigation of human beings, but also for the engineering of things used in daily life.

3 Investigation of Japanese sentences

In this work, we consider the process of sentence understanding as the analysis of the syntactic structure of a sentence, and we assume that those items which must be stored in short-term memory when understanding a sentence are bunsetsus whose modifiees are not determined. ("Bunsetsu" is a Japanese technical grammatical term. A bunsetsu is like a phrase in English, but it is a slightly smaller component. *Eki-de* "at the station" is a bunsetsu, and *sono*, which means "the" or "its," is also a bunsetsu. A bunsetsu is roughly a unit referring to an entity. So a bunsetsu is thought to be an appropriate unit of recognition.) Figure 1 is an example of calculating the number of bunsetsus whose modifiees are not determined in each step when analyzing the syntactic structure of the following sentence from left to right.

sono shounen-wa chiisai ningyou-wo motteiru.
(the) (boy) (small) (doll) (have)
The boy has a small doll.

[2] Note that the following descriptions are not directly related to Miller's 7 ± 2 theory, but to short-term memory. Lewis's work, "Magical number two or three," discussed linguistic features related to short-term memory [4]. The work discussed the number of center-embedded sentences and theorized that in English only one main clause sentence and one center-embedded sentence, for a total of two sentences, are allowed. In Japanese, one main clause sentence and two center-embedded sentences, for a total of three sentences, are allowed. These limitations are caused by the constraints of short-term memory, and have been discussed in English in principle four, "Two sentences", of Kimball's Seven Principles [2]. This research suggests that the reason for the limited number of center-embedded sentences is the limited capacity of human short-term memory.

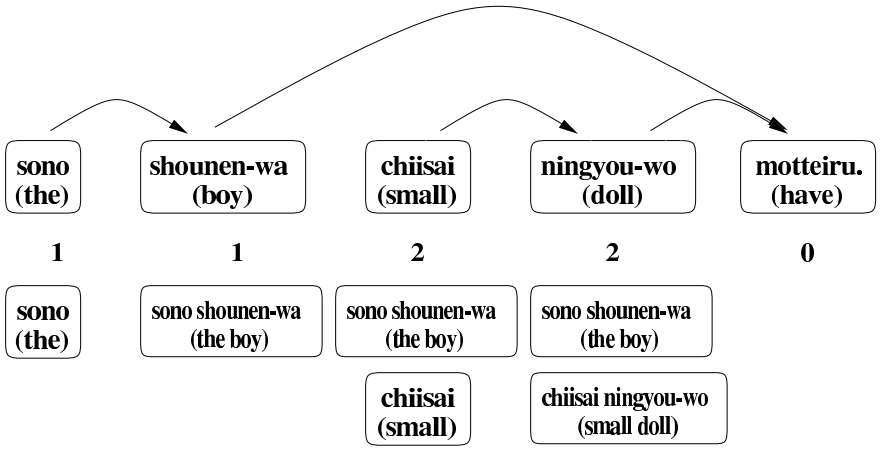

Fig. 1. How to estimate the number of bunsetsus whose modifiees are not determined

Arrows in the figure indicate the dependency structure. The number indicates the number of bunsetsus whose modifiees are not determined, and the lower part indicates the elements which must be stored in short-term memory. At the beginning, when *sono* (the) is input, its modifiee has not been determined yet, so it must be remembered. It is then stored in short-term memory as a bunsetsu whose modifiee is not determined. When *shounen* (boy) is input, *sono* (the) is found to modify *shounen* (boy). So *sono* (the) will not be used in the syntactic analysis after that, and it does not need to be remembered independently. *Sono* (the) is recognized to be attached to *shounen* (boy) in the form of *sono shounen* (the boy). As a result, only one element, *sono shounen* (the boy), whose modifiee is not determined, is stored in short-term memory. Next, *chiisai* (small) is input. This time, the dependency structure is not changed, and *sono shounen* (the boy) and *chiisai* (small) are stored in short-term memory. Next, when *ningyou* (doll) is input, *chiisai* (small) is recognized to modify *ningyou* (doll). *Chiisai* (small) will not be used in later analysis, because it is recognized to be attached to *ningyou* (doll) in the form of *chiisai ningyou* (small doll). Only the two elements *sono shounen* (the boy) and *chiisai ningyou* (small doll) are stored. Finally, *motteiru* (have) is input. Here, all the relationships of the dependency structure are determined and the number of bunsetsus with undetermined modifiees is 0. All the elements which were stored in short-term memory are cleared.

We assume that all human beings understand sentences the above way. The results are shown in Table 1. The number in the "bunsetsu" column is the number of bunsetsus having the given number of undetermined modifiees among all the bunsetsus of the Kyoto University corpus, (19,954 sentences and 192,352 bunsetsus). The number in the "sentence" column is the number of sentences having the given number of undetermined modifiees. In this table, only three

Table 1. Number of bunsetsus which undetermined modifiees

Bunsetsus with undetermined modifiees	Frequency	
	Bunsetsu	Sentence
0	19954	90
1	52751	1352
2	59494	5022
3	38465	6823
4	15802	4468
5	4488	1593
6	1143	480
7	195	102
8	47	17
9	10	5
10	3	2

bunsetsus exceeded the upper bound of Miller's 7 ± 2 rule. The result supports Miller's theory.

4 Investigation of English sentences

In the investigation of the Japanese corpus in the previous section we estimated the upper bound of the short-term memory required for sentence understanding. This section describes a similar investigation of an English corpus.

Yngve described a method for estimating the short-term memory capacity required in the syntactic analysis of an English sentence [8]. This method supposes that the nonterminal symbols, i.e., S and NP, which are stored in a stack when analyzing a sentence in a top-down fashion by using a push-down automaton, are those which need to be stored in short-term memory, and it counts the number of symbols stored in the stack. Figure 2 shows how the number of nonterminal symbols stored in a stack is counted in the analysis of the sentence, "The boy has a small doll," in a push-down automaton. Boxes in the lower part of Figure 2 indicate the state of the stack as the sentence is parsed. For example, at the beginning of the sentence, "The" is input first. When the sentence is analyzing in a top-down fashion, S is given first. Next, S is transformed into (NP VP). When VP is remembered, NP is transformed into (DT N). When N is remembered, DT is recognized to be "The"[3]. As a result, the two non-terminal symbols, VP

[3] Yngve's method has the following two problems. The first is, in Figure 2, we can select two possible patterns, (DT N) and (DT J N), in transforming NP, and we cannot select one of them when "The" is input. The other is that, by changing the grammar used in a corpus, the structure of a syntactic tree is changed and the result is changed. Despite these problems, we used Yngve's method because it is very easy to count with.

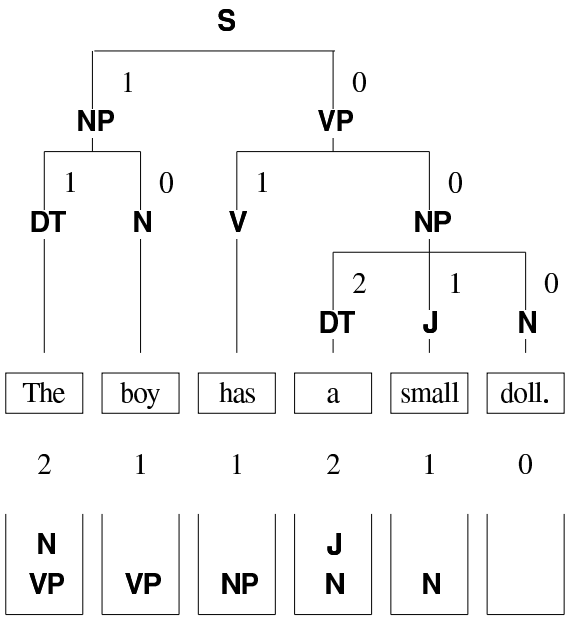

Fig. 2. How to count the number of nonterminal symbols stored in stacks

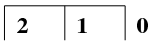

Fig. 3. How to give a number to each branch

and N, need to be stored in a stack while "The" is processed. Similarly, the nonterminal symbols which need to be stored for each stack are shown in Figure 2. The numbers of symbols in the stacks for each word are 2, 1, 1, 2, 1, and 0. Yngve also proposed an easy method of counting the number of nonterminal symbols stored in a given stack. In this method a number is assigned to each branch of a tree as shown in Figure 3. The sum of the numbers in the path from S to a word is considered as the number of symbols stored in a stack at that word. For example, at the word "The", "1, 1" is in the path of S, NP, DT, and "The", so the sum is 2, which matches the number of symbols stored in the stack.

Using this method, Sampson analyzed the SUSANNE corpus (130,000 words) and obtained the results shown in Table 2(a) [7]. "Frequency (words)" means the frequency of words with the corresponding number of nonterminals stored in a stack. With this method of analysis many sentences exceeded the upper bound of 7 ± 2, i.e., 9. Sampson counted again, changing the number of each branch, as in Figure 4. With this new method, when A is recognized, B, C, D, and E

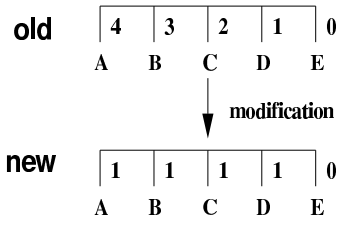

Fig. 4. Sampson's counting method

Table 2. Number of nonterminals stored in a stack (SUSANNE corpus)

(a)Yngve's method

Stack	Frequency (words)
0	7851
1	30798
2	34352
3	26459
4	16753
5	9463
6	4803
7	2125
8	863
9	313
10	119
11	32
12	4
13	1

(b)Sampson's method

Stack	Frequency (words)
0	55866
1	64552
2	12164
3	1274
4	76
5	4

are not remembered independently, but as one set of B, C, D, and E. Using this method, Sampson obtained the results shown in Table 2(b). This result showed that none of the sentences exceeded the lower bound of 7 ± 2, i.e., 5, therefore does not conflict with Miller's 7 ± 2 theory.

We followed the same methods in an analysis of the corpus of The Wall Street Journal of Penn Treebank [5]. We did not use the SUSANNE corpus because its structure is complicated, it is smaller than the Penn Treebank corpus, and it has already been studied by Sampson. The results for the Penn Treebank corpus are shown in Table 3. "Words" means the frequency of words having a given number of nonterminals stored in the stack. "Sentences" means the frequency of sentences having a given number of nonterminals stored in the stack. This time, we eliminated symbols such as periods, and we counted by changing the number of each branch in a coordination clause as in Figure 5, because the Penn

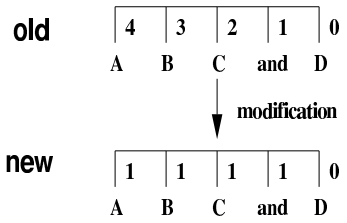

Fig. 5. How to modify the number of each branch in a coordination clause

Table 3. Number of nonterminals stored in a stack (Penn Treebank corpus)

Yngve's method in a word

Stack	Frequency	
	Words	Sentences
0	49208	132
1	377740	772
2	309255	3921
3	213294	9528
4	103864	13324
5	44274	11163
6	16478	6158
7	5750	2719
8	1939	981
9	661	338
10	243	111
11	92	29
12	43	17
13	15	14
14	1	1

Table 4. Number of nonterminals stored in a stack (Penn Treebank corpus)

Sampson's method in a word

Stack	Frequency	
	Words	Sentences
0	49208	132
1	485849	1956
2	414945	13367
3	140611	22966
4	28317	9124
5	3616	1518
6	283	133
7	28	12

Table 5. Number of nonterminals stored in a stack (Penn Treebank corpus)

Yngve's method in a NP

Stack	Frequency	
	NPs	Sentences
0	69820	4546
1	102337	7634
2	74126	16847
3	30025	11489
4	11432	5780
5	3336	2020
6	963	633
7	273	187
8	76	51
9	29	13
10	13	8

Treebank corpus is constructed such that the extra number of nonterminals in a coordination clause is counted. The results in Table 3 were found to match those in Table 2(a). Again, many sentences exceed the upper bound of seven plus or minus two. We also counted using Sampson's method. The results are shown in Table 4. Although the number of nonterminal symbols of the SUSANNE corpus did not exceed five, the Penn Treebank corpus included words with up to seven nonterminal symbols.

We also developed a new counting method for an English corpus which is different from Yngve's and Sampson's methods. Our method is based on an idea that we should not use, as a cognitive unit, words but phrases, which corresponds to bunsetsus, which are the units for counting in Japanese. We assume that human beings recognize NPs all at once instead of dividing them into words, and count the number of nonterminals stored in a stack at the NP level. In other words, we counted by using the sum of the numbers in the path from S to NP. The results shown in Table 5, are very similar to the results for Japanese sentences, shown in Table 1, and contain sentences with eight and nine NPs, which correspond to the plus-two part of Miller's 7 ± 2 theory. These results show our method to be effective.

Yngve's method did not obtain results that agree with Miller's 7 ± 2 theory, but Sampson's method and our method did. However, our method has the following two advantages over Sampson's method.

- Our counting method in English, which uses bunsetsu-corresponding NPs as the unit for counting, is based on our counting method for Japanese. (It is plausible for several languages to have the same level of cognitive units.)
- Although Sampson's method does not result in sentences with eight or nine nonterminal symbols, which is the upper bound of the 7 ± 2 theory, our method produced results that did. (Since "± 2" indicates an individual-based

variation, a method that does not result in sentences with eight or nine nonterminals for a large corpus is very unnatural.)

5 Conclusion

We investigated Miller's 7 ± 2 theory using Japanese and English corpora. New information obtained in this paper is shown here.

- When bunsetsus were used as the cognitive unit, the results of the investigation of Japanese syntactic recognition agreed with Miller's 7 ± 2 theory.
- When NPs were used as the cognitive unit, the results of the investigation of English syntactic recognition agreed with Miller's 7 ± 2 theory. This indicates that NPs are likely to be the cognitive unit. It seems natural that the NP level is the cognitive unit, because it is the same level as the Japanese cognitive unit, bunsetsu[4].
- If we suppose that bunsetsus and NPs are the cognitive units, the analyses in Japanese and English support Miller's 7 ± 2 theory and also support Yngve's theory [8], which is that the number of items stored in short-term memory does not exceed 7 ± 2 in language understanding and generation. These analyses support Miller's 7 ± 2 theory and Yngve's theory. From the standpoint of natural language processing, if Yngve's assertion is right, the assertion that "the number of items stored in short-term memory does not exceed 7 ± 2" can be used in the construction of an practical NLP system.

References

1. Charles J. Fillmore. The case for case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, Inc., 1968.
2. John Kimball. Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47, 1973.
3. Sadao Kurohashi and Makoto Nagao. Kyoto University text corpus project. pages 115–118, 1997. (in Japanese).
4. Richard L. Lewis. Interference in Short-Term Memory: The Magical Number Two (or Three) in Sentence Processing. *Psycholinguistic Research*, 25(1):93–115, 1996.
5. Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):310–330, 1993.
6. George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.
7. Geoffrey Sampson. Depth in English grammar. *Journal of Linguistics*, 33:131–151, 1997.
8. Victor H. Yngve. A Model and an Hypothesis for Language Structure. *the American Philosophical Society*, 104(5):444–466, 1960.

[4] A cognitive unit is thought to be a case element [1] or a unit taking the case element in the transformation process from short-term memory to the semantic network of long-term memory. So it seems natural that the cognitive unit is the same level of phrase in Japanese and English.