Why object clefts are easier to process than subject clefts in Japanese: Frequency or expectation?

Barış KAHRAMAN†, Atsushi SATO‡, Hajime ONO‡† and Hiromu SAKAI‡‡

†Faculty of Education, Çankakale Onsekiz Mart University, Anafartalar Campus C1-217, Çankakale, 17100 Turkey
‡Faculty of Science & Engineering, Kinki University 3-4-1 Kowakae, Higashi-Osaka, Osaka, 577-8502 Japan
‡‡Faculty of Education, Hiroshima University 1-1-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, 739-8524 Japan

Abstract Previous studies have shown that both frequency and expectation for upcoming structures play an important role on sentence processing. However, it is still unclear which one of these factors has a stronger impact on sentence processing. In order to explore the possible effects of frequency and expectation; we conducted a corpus analysis in Japanese in the current study. We first calculated the distribution and transitional probabilities of subject and object clefts, and then compared them with reading time data in our previous study. The results showed the number of subject clefts was higher than object clefts, whereas transitional probability of object clefts was higher than subject clefts at the embedded verb position. The results indicate that expectation can account for the processing difficulty of clefts in Japanese, whereas the simple frequency failed to explain the processing difficulty.

Keyword Expectation, Frequency, Corpus Analysis, Sentence processing, Japanese, Subject Clefts, Object Clefts

なぜ日本語目的語分裂文は主語分裂文よりも処理しやすいのか ～頻度と予期の観点からの考察～

カリフラマン バルシュ 佐藤 淳 小野 創 酒井 弘

あらまし文処理を扱った先行研究では「頻度」と後続する構造に対する「予期」が重要な役割を果たしていると言われている。しかし、これらの要因の中どちらの方がより強い影響力かもつかは明らかではない。本研究では、「頻度」と「予期」の影響について検討するために、日本語でコーパス分析を行い、主語分裂文と目的語分裂文の分布及び遷移的確率を調べた。結果、主語分裂文の頻度の方が目的語分裂文の頻度よりも高いのに対して、埋め込み動詞の位置における目的語分裂文の遷移的確率の方が主語分裂文よりも高いことがわかった。これらの結果を先行研究で観察された読み時間のデータと照らし合わせることで、日本語における分裂文処理の難しさを「予期」で捉えられると予期「頻度」では捉えられないことがわかった。

キーワード 予期、頻度、コーパス分析、文処理、日本語、主語分裂文、目的語分裂文

1. Introduction

Previous studies have pointed out that experience is one of the most important factors that govern the human sentence processing (e.g., Gennari & MacDonald, 2008, 2009 [1,2]; MacDonald, 1999 [3]; MacDonald & Christiansen [4]; Wells, et al., 2009 [5]). According to these studies, experience is shaped by distributional pattern of linguistic input we are exposed to. However, it is still not fully understood what kind of linguistic input forms our sentence comprehension system.

According to one view, the frequency of particular structure is one of the most important factors (e.g., Gennari & MacDonald, 2008, 2009 [1,2]; Mak, et al., 2002 [6]; Reali & Christiansen, 2007 [7]). People process some structures more easily than other structures, because they are more familiar with frequent structures. For example, in English, subject relative clauses (SRs) are easier to process than object relative clauses (ORs) (e.g., King & Just, 1991 [8]; Staub, 2010 [9]). Reali & Christiansen (2007) reported that SRs are more frequent than ORs. Moreover, they found that SRs occur more frequently with proper nouns while ORs frequently occur with pronouns. Taking these distributional patterns of SRs and ORs into consideration, Reali & Christiansen (2007) conducted a series of experiments, and showed that ORs were processed more easily than SRs when pronouns were used within relative clauses [7]. Similarly, previous studies have shown that SRs frequently occur with...
animate head-nouns, whereas ORs occur more frequently with inanimate head-nouns (e.g., Gennari & MacDonald, 2008 [1]; Mak, et al., 2002 [6]). These studies have shown that when ORs were presented with animate head-nouns, they were processed more easily. Taken together, these studies imply that the processing asymmetry between SRs and ORs stems from their distributional difference.

From a different viewpoint, it has been argued that distributional patterns of linguistic input affect people’s expectation for upcoming constituents and structures. According to this view, entropy (uncertainty) about upcoming structures, namely conditional probabilities of upcoming structures which are derived from combinations of linguistic input is decisive factor in the sentence processing (e.g., Hale, 2003, 2006 [10][11]. For example, Hale (2006) showed that conditional probability of SRs is higher than ORs at relative pronoun. In other words, SRs are more likely to be expected than ORs at relative pronoun in English [11]. Therefore, SRs are easier to process than ORs.

Overall, both frequency and expectation seem to play an important role in sentence processing, and account for the difficulty of sentence processing. However, it is still unclear which one of these factors has a stronger impact on sentence processing. In other words, we do not fully understand whether a particular structure is processed more easily because it is more frequent or it is easier to expect upcoming structures from linguistic input.

Most of the claims regarding the importance of frequency and expectation were made from European languages, whereas contributions from Asian languages are comparatively limited (e.g., Japanese: Sato, 2011 [12]; Chinese: Wu, et al., 2009 [13]; Korean: Yun, et al., 2010 [14]). The current study uses Japanese cleft sentences, and attempts to provide insights on how frequency and/or expectation relates to sentence processing, and how our experience is shaped. In the next section, we will explain why cleft sentences in Japanese were chosen as the target structure.

2. Why Japanese Clefts?

In Japanese, Sato (2011) conducted a corpus analysis and compared the distribution of SRs and ORs. The results showed that frequencies of SRs and ORs did not differ significantly. Based on these results, Sato (2011) argued that the simple frequency of SRs and ORs cannot account for the processing difficulty of ORs in Japanese [12]. In Chinese, Wu, et al. (2009) conducted a corpus analysis and self-paced reading experiments. However, unlike Japanese, their results were in line with the distributional patterns of SRs and ORs [13]. In Korean on the other hand, Yun, et al. (2010) argued that the conditional probabilities of SRs and ORs account for the processing difficulty of ORs.

In Japanese, relative clauses are not marked explicitly. In Chinese, on the other hand, explicit relative clauses marker (i.e. DE) is used. In the case of Korean, relative clause verb is marked by adnominal form of predicates. These cross-linguistic differences indicate that ambiguity of relative clauses might be higher in Japanese than in Chinese and Korean. Since the possibility of occurrence of a relative clause is quite higher in Chinese and Korean, the impact of the frequency or expectation might have been stronger in these languages. In Japanese, on the other hand, since the possibilities of occurrence of other structures are also high at relative clause verb, the possible influences of frequency or expectation on sentence processing might have been masked. In order to provide convincing evidence from Japanese, regarding the possible effects of frequency and expectation, the investigation of unambiguous structures would be more helpful. For this purpose, cleft sentences provide a good test case.

In Japanese, cleft sentences are very similar to relative clauses (see 1). Particularly their word orders are identical, but the embedded verb is marked with no-wa in cleft sentences. The particle no is taken as a complementizer or nominalizer, and wa is a binding particle (Hiraiwa & Ishihara, 2002 [15]). We assume that the complex of the particle no-wa is a cleft marker. Unlike relative clauses, structural ambiguity can be resolved at the embedded verb, due to the use of no-wa. Moreover, Kahraman, et al (2011) have already compared the reading times of subject clefts (SCs) and object clefts (OCs) as shown in (1) [16].

1. a. SC condition:

Kyonen sobo-o inaka-de kaihoushita-nowa shinseki-da.
Last year grandma-acc village-loc nursed-NOWA relative-cop
‘It is the relative who nursed my grandmother last year at the village.’

1. b. OC condition:

Kyonen sobo-ga inaka-de kaihoushita-nowa shinseki-da.
Last year grandma-nom village-loc nursed-NOWA relative-cop

1. Due to the space limitations, examples were simplified. See Kahraman (2011) for complete list of examples [17].
'It is the relative who my grandmother nursed last year at the village.'

The results showed that OCs were read faster than SCs at the embedded verb position. Kahraman, et al. (2011) argued that in Japanese, the proximity-based accounts such as Structural Distance Hypothesis (O’Grady, 2007) [18] and Dependency Locality Theory (Gibson, 1998) [19] cannot account for the processing difficulty of clefts and relative clauses at the same time [16]. Furthermore, Kahraman, et al. (2011) speculated that the frequency of SCs and OCs or expectation difference for upcoming constituents at the embedded verb position might have affected the results. Therefore, the investigation of distributional patterns of cleft sentences in Japanese would provide insights on how frequency and expectation are related to sentence processing.

In the current study, in order to explore the relation between the frequency, expectation and sentence processing, we conducted a corpus analysis. The results indicated that expectation can account for the difficulty of SCs and OCs or expectation difference for upcoming constituents at the embedded verb position might have affected the results. Therefore, the investigation of distributional patterns of cleft sentences in Japanese would provide insights on how frequency and expectation are related to sentence processing.

In the current study, in order to explore the relation between the frequency, expectation and sentence processing, we conducted a corpus analysis. The results indicated that expectation can account for the difficulty of SCs and OCs or expectation difference for upcoming constituents at the embedded verb position might have affected the results. Therefore, the investigation of distributional patterns of cleft sentences in Japanese would provide insights on how frequency and expectation are related to sentence processing.

In the current study, in order to explore the relation between the frequency, expectation and sentence processing, we conducted a corpus analysis. The results indicated that expectation can account for the difficulty of SCs and OCs or expectation difference for upcoming constituents at the embedded verb position might have affected the results. Therefore, the investigation of distributional patterns of cleft sentences in Japanese would provide insights on how frequency and expectation are related to sentence processing.

In the current study, in order to explore the relation between the frequency, expectation and sentence processing, we conducted a corpus analysis. The results indicated that expectation can account for the difficulty of SCs and OCs or expectation difference for upcoming constituents at the embedded verb position might have affected the results. Therefore, the investigation of distributional patterns of cleft sentences in Japanese would provide insights on how frequency and expectation are related to sentence processing.

3. Corpus Analysis

We first compared the simple frequency of SCs and OCs, and calculated their transitional probabilities at the embedded verb position.

In the current study, we used a part (3 million words) of the corpus KOTONOHA (10 million words from written Japanese developed by the National Institute for Japanese Language). Due to the absence of a large-scale parsed corpus in Japanese, we conducted an automated morpheme analysis by Mecab 0.98 (developed by Taku Kubo). In Japanese, cleft sentences are marked with the morpheme no-wa. In order to extract cleft sentences, we used ChaKi.NET (developed by NAIST), and selected them by pulling out NOWA. Extracted sentences were manually classified into SCs, OCs, other clefts and non-clefts. Examples are as follows:

(2) SCs:
Teki-o taoshita-nowa Salamanca-da-tta.
enemy-ace overthrew-NOWA Salamanca-cop-past
‘It was Salamanca who overthrew the enemy’

(3) OCs:
Ore-ga aitenishiteita-nowa hanzai-sha-da-tta.
I-nom deal with-NOWA criminal-cop-past
‘It was the criminal that I dealt with.’

(4) Other clefts:
Kaisatsuguchi-o deita-nowa 9:15-da.
ticket gate-acc got out-NOWA 9:15-cop.
‘It was 9:15 when I got out of the ticket gate.’

(5) Non-clefts
Tairyoku-ga atta-nowa iu-odemai
physical power-nom existed-NOWA needless to say
‘It is needless to say that there was a physical power.’

In SCs, the subject NP appears in the focused position, while in OCs, the object NP does so. In the case of other clefts, the focused element is adverbiaal phrases such as time, place, reason. In the case of non-clefts, although the verb is marked with no-wa, there is no copular, and the element appears after no-wa has a predicative relation with the original clause.

3.1. Results

3.1.1. Corpus Analysis 1: General Frequency

From our 3 million-word-corpus, 2085 sentences were extracted. We then manually counted these sentences. In total, the number of clefts was 1756 (84%), and the number of non-clefts was 329 (16%). Of these sentences, 656 sentences were tagged as SCs (31%) while 170 sentences were tagged as OCs (8%), and 930 sentences were tagged as other clefts (45%). A test of Chi-square showed that there was a significant difference among 4 types of sentences \( \chi^2 (3) = 662.97, p < .01 \). Ryan’s procedure showed that all of the 4 conditions significantly differed from each other (p < .01). This indicates that the number of SCs was statistically higher than OCs.

These results indicate that the distributions of SCs and OCs are inconsistent with the reading time data observed in Kahraman, et al. (2011) [16]. Although OCs were easier to process than SCs in the self-paced reading experiment, the number of SCs was higher than that of OCs in the corpora. Therefore, the processing asymmetry between SCs and OCs cannot be explained by their distributions.

However, in their reading experiment, Kahraman, et al. (2011) only used transitive verbs. In Corpus Analysis 1, we counted all kinds of cleft sentences regardless of their verb types. Therefore, the inconsistency between the reading time data and frequency may be due to the verb types. In order to test this possibility, we conducted Corpus Analysis 2 that takes verb types into consideration. Specifically, we compared the number of SCs and OCs

\[ \text{We included these sentences in the analysis, because they will be necessary while computing transitional probabilities.} \]
with transitive verbs which take nominative and accusative NPs as their arguments.

3.1.2. Corpus Analysis 2: Cleft sentences with Transitive Verbs

In total 752 sentences were extracted excluding passive, causative, ditransitive and intransitive verbs. Of these, 231 sentences were tagged as SCs (31%) while 131 sentences were tagged as OCs (17%). The number of other clefts was 296 (39%), and the number of non-clefts was 94 (13%). A test of Chi-square showed that there was a significant difference among 4 types of sentences. \( \chi^2 (3) = 136.16, p < .01 \). Ryan’s procedure again showed that all of the 4 conditions significantly differed from each other (p < .01). In other words, distributional tendencies of cleft sentences with transitive verb are very similar to their general tendencies in Corpus Analysis 1. This indicates that the inconsistency between the reading time data in Kahraman, et al. (2011) and frequency cannot be attributed to the verb types.

Overall, the results of the two analyses revealed that the simple frequencies of SCs and OCs cannot account for their processing asymmetry in Japanese. This suggests that other factors such as expectation for upcoming constituents might have a stronger impact on sentence processing. In the next subsection, we will attempt to explore some possible effects of expectation by calculating transitional probabilities of SCs and OCs at the embedded verb.

3.1.3. Transitional Probabilities

In order to calculate the processing difficulty of a particular structure, Hale (2003, 2006) proposed Entropy Reduction Hypothesis [10], [11]. According to Entropy Reduction Hypothesis, if the entropy, namely uncertainty about upcoming structures is greater, the processing becomes harder. When there are many possible continuations, entropy is high, and when the possibilities of upcoming continuations decrease, entropy is reduced.

In other words, if the predictability of an upcoming structure is higher than another structure, it is processed more easily.

In order to calculate the amount of entropy reduction, conditional entropies of possible structures should be computed. In order to do this, formalized grammar is necessary [10], [11]. However, unlike English, there is no fully parsed corpus in Japanese. Therefore, calculation of entropy reduction of a structure seems quite hard in Japanese. Nevertheless, Aslin, et al. (1998) pointed out that transitional probabilities are functionally equivalent to conditional probability statistics such as conditional entropy [20]. In other words, transitional probabilities and conditional entropy make similar predictions for the processing difficulty. Transitional probability is defined as a conditional probability measuring the predictability of adjacent elements (e.g, Aslin, et al., 1998 [20]; Pelucchi, et al., 2009 [21]; Saffran, et al., 1996 [22]; Thompson & Newport, 2007 [23]). The formulation of transitional probability is as shown in (6).

\[
\text{(6) probability of } Y|X = (\text{frequency of } XY) / (\text{frequency of } X)
\]

For example, if structure X is always followed by structure Y, transitional probability of Y at X is 1 (100%). On the other hand, if structure X is followed by structure Y or Z, and if the frequencies of Y and Z are equal, their transitional probabilities are 0.5. When transitional probability of a particular structure increases, its expectation becomes higher.

In the present analysis, we applied transitional probability formula to SCs and OCs with transitive verbs. In other words, we calculated transitional probabilities of SCs and OCs at the embedded verb position. In order to compute transitional probability of SCs, we divided the total frequency of SCs into total frequency of \([\text{accusative-NP + verb-NOWA}]\) sequence. In the case of OCs, total frequency of OCs was divided into total frequency of \([\text{nominative-NP + verb-NOWA}]\) sequence.\(^3\)

The results are as follows. Transitional probability of SCs within \([\text{accusative-NP + verb-NOWA}]\) sequence was .57 (205/357). On the other hand transitional probability of OCs within \([\text{nominative-NP + verb-NOWA}]\) sequence was .75 (86/114).\(^4\) The results showed that transitional probability of OCs within \([\text{nominative-NP + verb-NOWA}]\) sequence was higher than that of SCs within \([\text{accusative-NP + verb-NOWA}]\). These results indicate that although the frequency of OCs is lower than SCs, the Japanese parser’s certainty about OCs is higher than SCs. In other words, proportion of expectation for OCs at the embedded verb position is higher than that for SCs. If we assume that this kind of distributional pattern is learned and used for making predictions in the sentence processing, the processing asymmetry between OCs and SCs in Japanese can be explained by expectation. In the next section, we will

\(^3\) Sentences in which an element intervened between NP and verb were also included in the analysis.

\(^4\) The number of SCs and OCs differ from the Corpus Analysis 2, because we only included clefts in which subject or object explicitly expressed before the embedded verb.
discuss these results in more detail.

4. General Discussion

In the current study, we investigated into the question how the frequency and expectation are related to sentence processing, and how human experience is shaped. In order to answer these questions we investigated the frequency and transitional probabilities of subject and object clefts, and compared these results with previous processing data of Japanese clefts [16]. Main findings of the current study can be summarized as follows. The simple frequency of SCs was higher than that of OCs, whereas transitional probability of OCs was higher than that of SCs at the embedded verb position.

Reali and Christiansen (2007) showed that subject relative clauses (SRs) are more frequent than object relative clauses in English, and argued that the processing asymmetry between SRs and ORs can be attributed to their distribution [7]. In the case of Japanese clefts, the results are considerably different from Reali and Christiansen. Even though OCs were easier to process than SCs [16], SCs were more frequent than OCs. Therefore, the processing asymmetry between SCs and OCs in Japanese cannot be attributed to their distributional pattern in a corpus. Sato (2011) has also shown that the distributions of SRs and ORs are not reflected in their processing difficulty in Japanese [12].

Taken together our results suggest that the simple frequency alone cannot account for the processing difficulty of relative clauses and cleft sentences in Japanese (Roland et al., 2007 [24]).

While the simple frequencies of SCs and OCs are not in line with their processing difficulty, their transitional probabilities are consistent with the difficulty pattern observed in Kahraman et al (2011) [16]. At the embedded verb position, transitional probability of OCs was higher than that of SCs. In other words, the uncertainty about SCs was greater than OCs, indicating that the processing difficulty of SCs can be explained by expectation [10], [11], [14]. Our study indicates an important possibility that the effects of frequency observed in previous studies might be attributed to transitional probabilities. In other words, the distribution of noun types such as proper noun–pronoun or animate–inanimate might have changed expectation of upcoming constituents and this would have affected the results (e.g., Reali & Christiansen, 2007 [7]; Gennari & MacDonald, 2008, 2009 [1], [2]; Mak, et al., 2002 [6]; Wu, et al., 2009 [13]). However, in the present study we could not analyze the distribution of noun types within cleft sentences. In order to examine their possible effects we need to conduct more detailed analyses.

As we will discuss below, our results are share some similarities with child language acquisition studies. Previous studies in child language acquisition have reported that transitional probabilities have a strong impact on the acquisition of word segmentation by young children (e.g. Aslin, et al., 1998 [20]; Pelucchi, et al., 2009 [21]; Saffran, et al., 1996 [22]; Thompson & Newport, 2007 [23]). For example, Aslin et al. (1998) used an artificial language to test the possible effects of transitional probabilities and frequency in acquisition of word segmentation by 8-month-old children. They manipulated the frequency and transitional probability of words. The artificial words were matched in frequency, but differed in their transitional probabilities. The results showed that 8-month-old children used transitional probabilities to segment words, whereas they did not use the frequency information. This suggests that transitional probabilities are likely to play a more important role on the language acquisition than frequency does.

Although our study is not about learning per se, and the linguistic unit is quite different from child language acquisition studies, the results are in line with respect to stronger effects of transitional probabilities. This indicates in both language acquisition and adult sentence processing, expectation for upcoming continuations would have stronger impact than the simple frequencies. It thus can be said that transitional (conditional) probabilities are likely to play a more crucial role than the frequency in the formation of our experience and language comprehension system.

4.1. Limitations and Future Studies

As we discussed above, in the present study, we could not analyze the distribution of proper noun, pronouns, animate and inanimate nouns within cleft sentences. However, previous studies showed that these kinds of lexical items are also related to sentence processing (e.g., Gennari & MacDonald, 2008 [1]; Mak, et al., 2002 [6]; Wu, et al., 2009 [13]). In future studies we need to explore the distribution of these noun types and examine their possible influence on transitional probabilities of upcoming structures and sentence processing.

Another limitation of the current study is the gap between the corpus analyses and reading time data in Kahraman, et al. (2011) study. In the current study, we did not manipulate the test sentences nor run any experiment, based on the distributional patterns of cleft
sentences. In order to draw more conclusive conclusions regarding to effects of probabilistic factors such as frequency and expectation, we need to conduct follow up experiments. We leave these issues for future studies.

5. Conclusions
In order to explore possible effects of frequency and expectation on sentence processing, we conducted corpus analyses, and compared the distribution and transitional probabilities of subject and object clefts in Japanese. The results showed that the transitional probabilities successfully account for the processing difficulty of SCs, whereas the simple frequency cannot, indicating that expectation has a stronger impact than the simple frequency on sentence processing. Overall, our results suggest that, to integrate the sentence processing and probabilistic factors, we need to investigate corpora from different dimensions (Roland, et al., 2007 [24]).

References

Acknowledgments
This research was supported by (1) Grant-in-Aid for Scientific Research (B) “Neurocognitive basis for language learning through the processing of input and output (PI: Hiromu Sakai, #20320060)” by JSPS; (2) Grant-in-Aid for Scientific Research (B) (PI: Kentaro Nakatani, #21320083) by JSPS; (3) Grant-in Aid for Young Scientists from the MEXT (PI: Hajime Ono, #21720152). We would like to thank Rosalynn Chiu for her help. All remaining errors are our own.