Predicting Japanese scrambling in the wild

Naho Orita
Graduate School of Information Sciences
Tohoku University
naho@ecei.tohoku.ac.jp

Abstract

Japanese speakers have a choice between canonical SOV and scrambled OSV word order to express the same meaning. Although previous experiments examine the influence of one or two factors for scrambling in a controlled setting, it is not yet known what kinds of multiple effects contribute to scrambling. This study uses naturally distributed data to test the multiple effects on scrambling simultaneously. A regression analysis replicates the NP length effect and suggests the influence of noun types, but it provides no evidence for syntactic priming, given-new ordering, and the animacy effect. These findings only show evidence for sentence-internal factors, but we find no evidence that discourse level factors play a role.

1 Introduction

Speakers constantly make choices about the form of their utterances such as referring expressions (Givón, 1983; Ariel, 1990; Gundel et al., 1993) and word order (Bock and Irwin, 1980; Arnold et al., 2000; Birner and Ward, 2009). For example, Japanese speakers have a choice between canonical SOV and scrambled OSV to convey the same meaning of the sentence as in (1).

(1) a. Taro-ga inu-o oikaketa.
Taro-NOM dog-ACC chased
'Taro chased the dog.'

b. Inu-o Taro-ga oikaketa.
dog-ACC Taro-NOM chased
'Taro chased the dog.'

The positioning of the direct object in (1b), scrambling (Saito, 1985; Saito and Hoji, 1983; Miyagawa, 1997), is known to be sensitive to such factors as length of the noun phrase (Yamashita and Chang, 2001), structural priming (Yamashita et al., 2002), given-new ordering of discourse participants (Ferreira and Yoshita, 2003), and animacy (Tanaka et al., 2011).

These experiments in language production examine the influence of one or two factors in a controlled experimental setting with a set of homogeneous stimuli that typically occur infrequently in the real world. However, as intensively discussed in Jaeger (2010), speakers' choices of sentence structure rather depend on the presence and interaction of multiple factors that are influenced by the probability distribution of preceding inputs.

This study complements previous psycholinguistic experiments in Japanese scrambling. We use a more representative set of Japanese sentences than those used in previous language production experiments and simultaneously test multiple effects on scrambling using a single regression analysis. This effort constitutes an initial step toward understanding Japanese scrambling in the wild that goes beyond laboratory data. The analysis replicates the effect of noun phrase length and suggests the influence of noun types, but it provides no evidence for syntactic priming, given-new ordering, and the animacy effect.

2 Related Work

2.1 Language production experiments

Although a body of research concerns the comprehension of scrambled sentences in Japanese (Koizumi and Tamaoka, 2010; Koizumi and Imamura, 2016, among many), the production counterpart has not been investigated to the same extent. Previous findings suggest that (i) Japanese speakers tend to put a longer object before a short subject via scrambling (Yamashita and Chang, 2001), (ii)

Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017), pages 41–45,
Valencia, Spain, April 7,

tend to produce the same structure as the prime sentence (Yamashita et al., 2002), (iii) are more likely to position animate entities earlier in the sentence than inanimate entities independent of their grammatical role and assign the subject role to animate entities (Tanaka et al., 2011), and (iv) tend to produce given arguments before new, where this effect is stronger when the previous mention of the given argument is lexically identical (Ferreira and Yoshita, 2003).

These experiments examine one or two hypotheses in a controlled setting to capture a precise mechanism of language production. However, it remains unknown whether and to what extent multiple effects contribute to observable behavior. This study complements these works by simultaneously testing multiple effects with naturally distributed data.

2.2 Corpus studies

Corpus studies examine the effect of some of the factors suggested above with more natural linguistic data. In Yamashita (2002), 19 scrambled sentences are found out of 2,635 sentences in Japanese magazines. Of these 19 scrambled sentences, 14 involve scrambling of "heavy" constituents that contain relative or subordinate clauses and 5 sentences contain anaphoric expressions that refer to preceding entities.

Imamura (2016) measures salience and information decay to investigate the distribution of subjects and objects in both SOV and OSV. He counts how many sentences stand between the argument and its referent to measure salience of the referent and how many times the argument is referred to after the target sentence to measure information decay. Analysis of 100 sentences for each word order that are randomly extracted from contemporary written Japanese texts (Maekawa et al., 2014) shows that the referent of the scrambled object tends to occur close to the scrambled sentence (i.e., old information), but is mentioned less frequently than the subject in successive sentences (i.e., not topical).

Although they explore more natural data than previous production experiments, they do not analyze the influence of multiple factors. Crucially, the sample size also is either limited (19 scrambled sentences in Yamashita (2002)) or intentionally balanced (100 sentences for each word order in Imamura (2016)). Moreover, they hand-code relevant information. These shortcomings make the analysis harder to extend to the wider data.

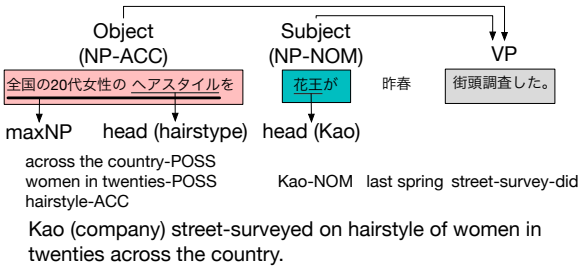

Figure 1: Illustration of the scrambled sentence extracted from the corpus

This study scales up these corpus studies by exploiting existing corpus annotations and by using a regression model that simultaneously tests multiple factors while handing unbalanced data. The following section describes the corpus and predictors in the analysis.

3 Predicting scrambling in Japanese

3.1 Corpus

We use NAIST Text Corpus (Iida et al., 2007) containing 2,929 documents (38,384 sentences) from Japanese newspapers. The corpus includes annotations of predicate-argument structure, event nouns, and coreferences along with morphological and syntactic information.

We use annotated dependency relations between arguments and the predicate to extract scrambled and canonical word order sentences. We extract pairs of nominative (NP with case marker *-ga*) and accusative (NP with case marker *-o*) arguments that depend on the same predicate (active voice only) as illustrated in Figure 1. We then check the linear order of these arguments. The result is 185 scrambled sentences and 2,918 canonical word order sentences. We exclude null arguments, which occur frequently in Japanese. The relatively small number of extracted sentences is primarily attributable to this exclusion.

3.2 Predictors

Based on previous findings as in Section 2, we extract the following information from the corpus and include it in the analysis.

Syntactic priming: We approximate the syntactic priming effect (Yamashita et al., 2002) using equation (1). It decays exponentially with the distance between the target sentence s_i and the

Predictor		$\hat{\beta}$	SE ($\hat{\beta}$)	z	p	$\chi^2(df)$	p_χ
SYNTACTIC PRIMING		0.67	0.57	1.16	0.24	1.23 (1)	0.27
SUBJECT GIVEN-NEW		-0.12	0.22	-0.53	0.60	0.28 (1)	0.60
OBJECT GIVEN-NEW		0.15	0.22	0.69	0.49	0.47 (1)	0.49
SUBJECT LENGTH		-0.15	0.02	-6.51	7.76e-11	65.28 (1)	6.49e-16
OBJECT LENGTH		0.12	0.01	10.59	<2e-16	107.71 (1)	3.11e-25
SUBJECT NOUNTYPE	NAME	-0.69	0.21	-3.26	0.001	16.70 (3)	<0.001
	VERBAL	-1.19	0.53	-2.26	0.02		
	FORMAL	-0.48	0.90	-0.54	0.59		
OBJECT NOUNTYPE	NAME	0.13	0.53	0.25	0.81	97.35 (3)	5.77e-21
	VERBAL	-0.36	0.20	-1.83	0.07		
	FORMAL	3.36	0.34	9.80	2e-16		

Table 1: Logistic model predicting scrambling for each sentence

scrambled sentence s_j as previously mentioned. If there is no scrambled sentence in the preceding discourse, we use parameter α that represents how likely speakers produce a scrambled sentence. We set this value at 0.01 in the analysis.

$$f(d_{i,j}) = \begin{cases} e^{-d_{i,j}/a} & \text{if } s_j \text{ exists} \\ \alpha & \text{if no } s_j \end{cases} \quad (1)$$

Given-new ordering: We measure the effect of given-new ordering, in particular, the effect of lexically identical mentions (Ferreira and Yoshita, 2003) by using the same function in (1). If there is a lexically identical word in the preceding discourse, we compute the distance between the target sentence s_i and the sentence containing the lexically identical word. We check all content words in the maxNP (a maximally spanning noun phrase as in Figure 1) and use the closest previous mention for computation. This is not a precise representation of given-new ordering in that we do not use coreference relations in contrast to Imamura (2016) that hand-annotated coreferences including bridging references.[1]

Length: The length of subject and object maxNPs in letters are included because the length has been shown to affect scrambling in Japanese (Yamashita and Chang, 2001; Yamashita, 2002).

Noun type: Scrambling is also correlated with animacy (Tanaka et al., 2011). There is no available animacy annotation in Japanese, but nouns in the NAIST Text Corpus are annotated with the type of information, such as proper names, verbal nouns, and formal nouns. Verbal nouns such as *kettei* 'decision' and *suisen* 'recommendation' can be verbs with the addition of a light verb *-suru* 'do'. Formal nouns such as *-koto* and *-mono* are a kind of nominalizer that lacks semantic content. Some formal nouns such as *-koto* and *-no* often function as a complementizer that takes a clausal complement (Inoue, 1976). We include these noun types in the analysis to capture automatically a coarse distinction between animate nouns and event/clausal expressions. Only the types of head nouns (illustrated in Figure 1) are included in the analysis. Other noun types such as common nouns (e.g., 'cat' and 'society') and temporal nouns (e.g., 'tomorrow') are excluded, but they account for the majority of the nouns in the corpus (about 65%).

4 Results

We use a logistic regression model wherein the outcome variable is the word order —i.e., scrambled or canonical. We include all predictors described above. The model was fitted using `glm` in R. Table 1 summarizes coefficient estimates $\hat{\beta}$, standard errors SE($\hat{\beta}$), associated Wald's z-score, and the significance level p for the predictors described above. Positive (negative) coefficients indicate a higher (lower) probability of scrambling. We also report the chi-square value that indicates how much the model is improved by including the predictor. We use the variance inflation factor to assess multicollinearity. Scores range between 1.01 and 1.19, suggesting non-significant influence on the reliability of the parameter estimates (Neter et al., 1996).

Syntactic priming (mean = 0.03, sd = 0.12) fails to achieve significance, primarily because of the distribution of scrambled sentences: only two occur in the same document.[2]

Given-new ordering does not attain significance (subject: mean = 0.42, sd = 0.40, object: mean

[1]The NAIST Text corpus includes coreference annotation, but there are only 8 scrambled objects that are coreference-annotated. On the other hand, there are 92 scrambled objects that include lexically identical mentions (old information).

[2]Including scrambled sentences with topicalized objects (case marker *-wa*) in the analysis did not boost syntactic priming, as only 112 such sentences occur in the corpus.

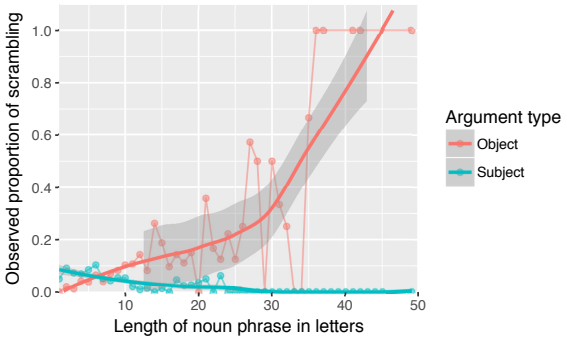

Figure 2: Observed proportion of scrambling by NP length in letters: bold lines are smoothed trend lines with confidence interval around.

= 0.3, sd = 0.38), probably because there is no contrast between scrambled objects and scrambled subjects/canonical objects. [3]

The significant negative correlation with the length of subject NP and the positive correlation with the length of object NP replicate previous studies. The empirical distribution of the proportion of scrambling by the length of subjects and objects in Figure 2 confirms this tendency.

Noun type highly influences the choice of scrambling. A negative correlation of SUBJECT-VERBAL seemingly conflicts with the animacy effect at first glance, but post-hoc examination of the distribution of noun types in Table 2 shows that SUBJECT-VERBAL tends to occur with OBJECT-VERBAL, compared to SUBJECT-FORMAL and SUBJECT-NAME. This suggests there is less inclined to scrambling when both are verbal nouns. SUBJECT-NAME is a significant predictor that correlates negatively with scrambling, showing a strong relation between animate entities and subjects in the canonical position. However, the absence of a positive correlation of OBJECT-NAME indicates the animacy effect is not replicated.

The significant positive correlation of OBJECT-FORMAL suggests that speakers might prefer to place the clausal complement earlier in the sentence. Table 2 shows a higher proportion of scrambling with OBJECT-FORMAL (28 out of 55). Of these 28 scrambled cases, 26 include the clausal objects (71%), whereas 14 of 27 canonical cases include them (51%).

[3] A Welch two-sample t-test: scrambled subjects vs. scrambled objects: $p = 0.66$, canonical objects vs. scrambled objects: $p = 0.33$.

	OBJ-NAME	OBJ-VERBAL	OBJ-FORMAL	OBJ-ELSE	Total
SUBJ-NAME	57 (1)	283 (8)	10 (3)	475 (23)	825 (35)
SUBJ-VERBAL	3 (0)	59 (0)	1 (0)	129 (4)	192 (4)
SUBJ-FORMAL	0 (0)	9 (0)	1 (1)	20 (1)	30 (2)
SUBJ-ELSE	36 (3)	620 (34)	43 (24)	1357 (83)	2056 (144)
Total	96 (4)	971 (42)	55 (28)	1981 (111)	3103 (185)

Table 2: Confusion matrix of noun types: values in the brackets indicate the number of times scrambling occurs.

5 Discussion

We tested multiple factors said to affect scrambling using the naturally distributed data. Although the analysis did not support syntactic priming, given-new ordering, and the animacy effect, we replicated the length effect and established the influence of noun types. These findings only show evidence for sentence-internal factors, but we found no evidence for discourse level factors.

The absence of discourse level effects may be due to the small sample size of scrambled sentences and the low rate of priming in a preceding discourse. It will be important to replicate our results on more extensive data to confirm our observed effects. To scale up the analysis, we will collect a large number of sentences from the web and extract those that are reliably parsed using the method suggested in Sasano and Okumura (2016). The lack of given-new ordering effect is presumably due to our naive procedure for estimating coreferences. For better estimation, future work will involve heuristics known to be robust in Japanese anaphora resolution tasks.

The effect of formal noun implies Japanese speakers' preference for putting the clausal complement earlier (Yamashita and Chang, 2001). The influence of noun type poses a question about what accounts for this *richer content earlier* preference. Future work will explore the potential causes that determine "richness" of the noun phrase.

Acknowledgments

We thank Ryohei Sasano, Tohoku communication science lab semantics group, and four anonymous reviewers for helpful comments and discussion.

References

Mira Ariel. 1990. *Accessing noun-phrase antecedents*. Routledge.

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, pages 28–55.

Betty J. Birner and Gregory Ward. 2009. Information structure and syntactic structure. *Language and Linguistics Compass*, 3(4):1167–1187.

J. Kathryn Bock and David E. Irwin. 1980. Syntactic effects of information availability in sentence production. *Journal of verbal learning and verbal behavior*, 19(4):467–484.

Victor S. Ferreira and Hiromi Yoshita. 2003. Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of psycholinguistic research*, 32(6):669–692.

Talmy Givón. 1983. *Topic continuity in discourse: A. quantitative cross-language study*, volume 3. John Benjamins Publishing.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139. Association for Computational Linguistics.

Satoshi Imamura. 2016. A. corpus based analysis of scrambling in Japanese in terms of anaphoric and cataphoric co-referencing. Ms., University of Oxford.

Kazuko Inoue. 1976. *Henkei bunpō to Nihongo (Transformational Grammar and Japanese)*. Taishūkan, New York.

Florian T. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Masatoshi Koizumi and Satoshi Imamura. 2016. Interaction between syntactic structure and information structure in the processing of a head-final language. *Journal of psycholinguistic research*, pages 1–14.

Masatoshi Koizumi and Katsuo Tamaoka. 2010. Psycholinguistic evidence for the VP-internal subject position in Japanese. *Linguistic inquiry*, 41(4):663–680.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Shigeru Miyagawa. 1997. Against optional scrambling. *Linguistic inquiry*, pages 1–25.

John Neter, Michael Kutner, William Wasserman, and Christopher Nachtsheim. 1996. *Applied linear regression models*. McGraw-Hill.

Mamoru Saito and Hajime Hoji. 1983. Weak crossover and move α in Japanese. *Natural Language & Linguistic Theory*, 1(2):245–259.

Mamoru Saito. 1985. *Some asymmetries in Japanese and their theoretical implications*. Ph.D. thesis, NA Cambridge.

Ryohei Sasano and Manabu Okumura. 2016. A. corpus-based analysis of canonical word order of Japanese double object constructions. In *Proceedings of the 54th annual meeting on Association for Computational Linguistics*.

Mikihiro N. Tanaka, Holly P. Branigan, Janet F. McLean, and Martin J. Pickering. 2011. Conceptual influences on word order and voice in sentence production: Evidence from Japanese. *Journal of Memory and Language*, 65(3):318–330.

Hiroko Yamashita and Franklin Chang. 2001. "Long before short" preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.

Hiroko Yamashita, Franklin Chang, and Yuki Hirose. 2002. Separating functions and positions: Evidence from structural priming in Japanese. In *15th CUNY Conference on Human Sentence Processing*.

Hiroko Yamashita. 2002. Scrambled sentences in Japanese: Linguistic properties and motivations for production. *TEXT & Talk: An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 22(4):597–634.