Empirical Verification of Meaning-Game-based Generalization of Centering Theory with Large Japanese Corpus

Shun SHIRAMATSU[†] **Kazunori KOMATANI**[†] **Takashi MIYATA**[††]
{siramatu, komatani}@kuis.kyoto-u.ac.jp miyata.t@aist.go.jp
Kôiti HASIDA[†††, ††] **Hiroshi G. OKUNO**[†]
hasida.k@aist.go.jp okuno@kuis.kyoto-u.ac.jp

†Graduate School
of Informatics,
Kyoto University
Yoshida-Honmachi,
Sakyo-ku,
Kyoto 606-8501, Japan

††CREST,
Japan Science and
Technology Agency (JST)
Akihabara Dai-Building, Soto-Kanda 1-18-13, Chiyoda-ku,
Tokyo 101-0021, Japan

†††ITRI, National Institute
of Advanced Industrial
Science and Technology (AIST)
Akihabara Dai-Building, Soto-Kanda 1-18-13, Chiyoda-ku,
Tokyo 101-0021, Japan

Abstract

Centering theory (Grosz et al., 1995) tries to explain relations among attention, anaphora, and cohesion. It has two theoretical limitations. The first is the lack of a principle behind these discourse phenomena. The second is that the *salience* of discourse entities has not been quantitatively defined, although it plays a critical role in this theory. Hasida et al. (1995, 1996) propose the *meaning game* as a more principled model of intentional communication based on game theory, and claim that it can derive centering theory. This claim, however, has not yet been verified on the basis of substantial linguistic data. In this paper, we formulate salience as a measurable quantity in terms of a *reference probability*. We also formulate preferences subsuming centering theory under this quantitative formulation of salience. The preferences are derived from the meaning game and entail more general predictions than those of conventional centering theory. These formulations overcome the above limitations of centering theory. By following them, we empirically verify our generalization with a large Japanese corpus. The experimental results show that there is positive correlation between the salience (reference probability) of an entity and the simplicity (utility) of a noun phrase which refers to the entity. They also indicate correspondence between the values of expected utility and the ranking of the transition states. These results indicate that our generalization is appropriate.

1. Introduction

Principled and quantitative modeling of discourse is important for analyzing and generating discourse. Centering theory (CT) is a model of discourse structures. It explains the relations among attention, anaphora, and cohesion (Iida, 1997). However, CT has had two theoretical limitations. The first is the lack of a general principle behind the discourse phenomena. Although some studies on CT have focused on analyzing surficial linguistic features without general principles, we consider that the principle of discourse phenomena must be addressed based on measurable quantities. The second is that “salience”, which plays a critical role in CT, cannot be verified based on large linguistic data because it is not formulated as a measurable quantity, but as heuristic rules.

We have investigated the general principle of CT. We adopted the meaning game (MG) (Hasida et al., 1995, 1996) framework because it gives a more principled explanation of the discourse phenomena than CT does. MG is a model of intentional communication (e.g., anaphora) based on game theory. Game players in game theory correspond to interlocutors in MG, and they decide their intentions and interpretations at the Pareto-optimum. Although Hasida et al. (1995) claimed that CT could be derived from the MG by formulating salience in terms of a reference probability, their claim has yet to be verified on the basis of substantial linguistic data. In this paper, we formulate the MG-based generalization of CT and verify it with a large corpus of Japanese newspaper articles. Furthermore, we quantitatively define salience by using multiple regression with a corpus for the MG-based generalization and for its verification.

2. Centering Theory and Its Two Issues

2.1 Centering Theory

In CT, a discourse is represented as a sequence of utterances [U_1, U_2, … ,U_n]. The "*center*" is a discourse entity which draws attention. The center is likely to be pronominalised. The "*salience*" represents the degree of attention to a discourse entity. The salience also represents the likelihood of pronominalization. The salience has been defined as a heuristic ranking in previous studies (see Section 2.2). Centers are categorized as follows:

- $Cb(U_i)$: The backward-looking center of the utterance U_i, which denotes the most salient discourse entity referenced in both the previous context and the current utterance U_i.
- $Cf(U_i)$: The forward-looking centers of U_i, which denote a list of entities sorted by their salience.
- $Cp(U_i)$: The preferred center of U_i, which is the most salient discourse entity in $Cf(U_i)$.

CT embodies as the following rules (preferences) based on the heuristics definition of salience.

- **Rule 1 (pronominalization)**: If any element in $Cf(U_i)$ is pronominalized, the $Cb(U_i)$ is also pronominalized.
- **Rule 2 (topic continuity)**: The transition states of centers between utterances (Table 1) are preferred in the following order: CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT.

Table 1: Transition states of centers between utterances

	$Cb(U_i) = Cb(U_{i-1})$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i) \neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

Rule 1 means that pronouns are more likely to refer to *Cb* than non-pronouns. Rule 2 represents the preference order among transition states according to the strength of topic continuity.

2.2 Two Issues

Conventional CT studies face two limitations:

1. Lack of principles behind the rules. CT does not explain why the two rules occur in discourse phenomena.
2. Salience is formalized neither objectively nor quantitatively, but heuristically (e.g., Cf-ranking). Such ranking is non-falsifiable (unscientific) and cannot be verified against real linguistic data.

The first limitation means that CT should have a hypothesis about the mechanisms behind discourse phenomena. The second limitation means that CT should be based on the quantitative definition of salience. Salience in CT is approximated by a heuristic ranking, called "*Cf-ranking*" (Walker et al., 1994), as follows:

English Cf-ranking: subject > object > indirect object > complement > adjunct
Japanese Cf-ranking: topic (zero or grammatical) > subject > indirect object > object > others

The above Cf-ranking depends on only grammatical function. While Strube et al. (1999) proposed an extended Cf-ranking integrated with information status and Nariyama (2001) proposed an extended ranking integrated with contextual information, these rankings are based on surficial observations without sufficient theoretical grounds. Although Poesio et al. (2004) discussed the parameters settings in CT, their discussion was also based on heuristic ranking.

Besides this second limitation, we also note that heuristic ranking is difficult to integrate with other features that influence salience (e.g., distance between the current utterance and the latest expression referring to the target entity).

We address the above two issues in the following sections.

3. Generalization of Centering Theory based on the Meaning Game

The meaning game (MG) is a hypothesis about a model of intentional communication based on game theory (Hasida, 1996). We adopted MG to give CT a general principle. The MG-based account of anaphora is a more principled hypothesis than that of CT, because MG is based on the general principle of decision-making. In MG, the interlocutors' expected utility is represented as:

$$\sum_{w \text{ refers to } e} Pr(e)Ut(w).$$

Here, the $Pr(e)$ is the *reference probability* of a discourse entity e, which is the probability that e will be referenced in the next utterance. $Ut(w)$ is the utility of expression w that refers to e. The lower the cost of speaking or hearing w is, the higher $Ut(w)$ becomes. Here, we assume that the value of $Pr(e)$ is shared by interlocutors. Under this assumption, the solution which provides the maximum expected utility is the interlocutor's Pareto-optimum because the expected utility is shared by them. We leave miscommunication out of consideration under that assumption.

Hasida et al. (1995) suggested that Rule 1 and 2 of CT can be derived from MG only in a few particular cases.

3.1 Derivation of Preference 1a and 1b

Rule 1 of CT is a preference about pronominalization. Hasida et al. derived Rule 1 from MG in the following case which involves little semantic bias, "he" tends to refer to "Fred", and "the man" to "Max".

U_1: Fred scolded Max.
U_2: He was angry with the man.

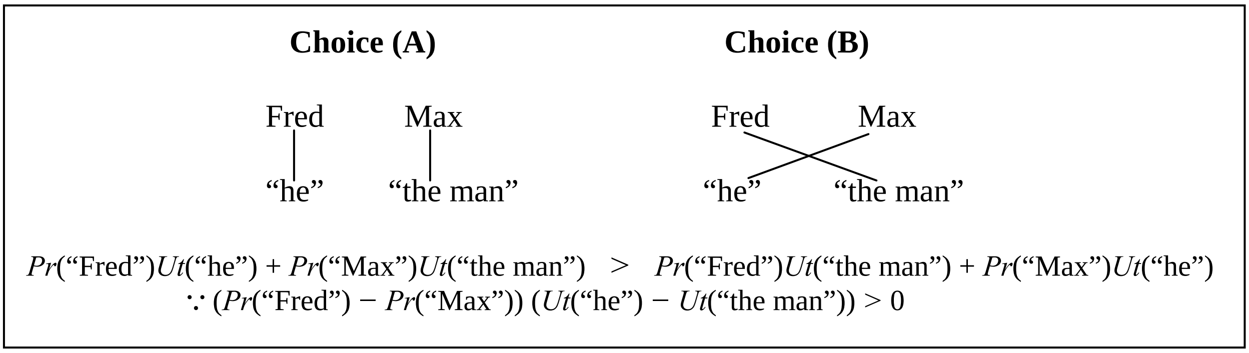
They assumed the following inequations in this case.

$Pr(\text{“Fred”}) > Pr(\text{“Max”})$ (∵ A subject is more salient than an object)
$Ut(\text{“he”}) > Ut(\text{“the man”})$ (∵ A pronoun costs less than a non-pronoun)

In this case, the interlocutors have two choices of anaphora. Hasida et al. indicated the above

Table 2: Correspondence between MG and CT concepts

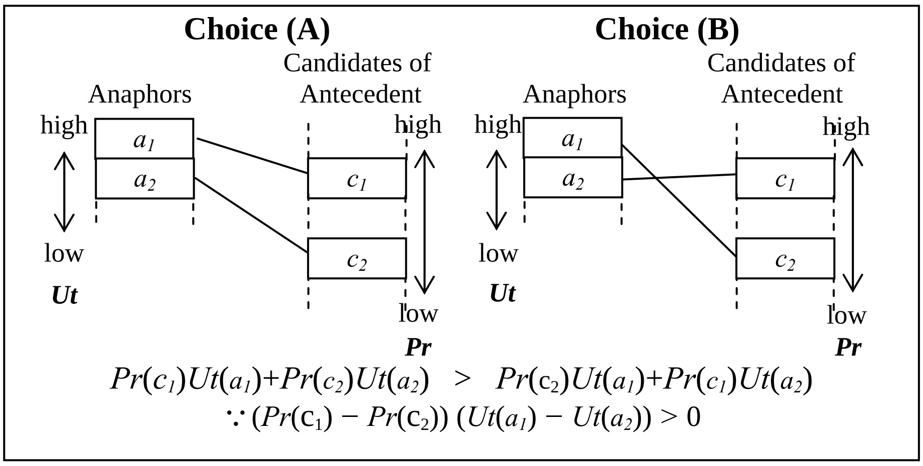
MG	CT
Pr: Reference probability	Salience (Cf-ranking)
High-*Pr* discourse entity	Center
Ut: Utility of noun phrase	Simplicity of noun phrase
High-*Ut* noun phrase	Pronoun
Low-*Ut* noun phrase	Non-pronoun



$$Pr(\text{“Fred”})Ut(\text{“he”}) + Pr(\text{“Max”})Ut(\text{“the man”}) \;>\; Pr(\text{“Fred”})Ut(\text{“the man”}) + Pr(\text{“Max”})Ut(\text{“he”})$$

$$\because (Pr(\text{“Fred”}) - Pr(\text{“Max”}))\,(Ut(\text{“he”}) - Ut(\text{“the man”})) > 0$$

Figure 1: Comparison of the expected utilities of two choices



$$Pr(c_1)Ut(a_1)+Pr(c_2)Ut(a_2) \;>\; Pr(\mathrm{c}_2)Ut(a_1)+Pr(c_1)Ut(a_2)$$

$$\because (Pr(\mathrm{c}_1) - Pr(\mathrm{c}_2))\,(Ut(a_1) - Ut(a_2)) > 0$$

Figure 2: Preference 1a

semantic bias by comparing the expected utilities of the two choices. In other words, a solution of their MG model is that choice (A) is preferred over choice (B) in Figure 1. This solution conforms to a prediction using Rule 1. Thus, they claimed this thought experiment proves that rule 1 of CT can be derived from MG. Table 2 shows correspondence between MG and CT concepts.

The above derivation, however, has neither been given a general formulation nor been verified on the basis of substantial linguistic data. An utterance in general examples possibly has a few anaphors and a lot of candidates of antecedent, whereas the above example case has only two anaphors and only two candidates of antecedent. Thus, general formulation is required before one can apply this model to real linguistic data. We generalize the above derivation to the following preference.

Preference 1a: When an utterance has multiple anaphors, an anaphor with a higher utility among them tends to refer to an entity with a higher reference probability.

Figure 2 illustrates Preference 1a we propose. In this example, choice (A) is preferred over choice (B). This is the preference in the cases of multiple anaphors in an utterance. Below, we generalize it to cases that do not depend on the number of anaphors in an utterance as follows:

Preference 1b: There is a positive correlation between the utility and the reference probability.

These preferences are based on a general principle. Moreover, the coverage of these preferences is wider and more general than that of Rule 1 of CT. Accordingly, these preferences we propose are generalizations of Rule 1.

3.2 Derivation of Preference 2

Rule 2 of CT is a preference about local cohesion that indicates the level of topic continuity. Transition states are categorized into four types with respect to two conditions (Table 1). These four types have been heuristically ranked by local cohesion or topic continuity. The first condition, $Cb(U_i) = Cb(U_{i-1})$, means that the current utterance U_i inherits Cb from the previous utterance U_{i-1}. This condition corresponds to cohesion between U_{i-1} and U_i. The second condition, $Cb(U_i) = Cp(U_i)$, means that $Cb(U_i)$ is the most salient entity in U_i. This condition corresponds to the prediction of cohesion between U_i and U_{i+1} because $Cp(U_i)$, the most salient entity in U_i, is the most likely one to be pronominalized in the following utterance U_{i+1}.

We consider that the preference order of Rule 2 is attributed to expected utility. When the first condition holds, the reference probability of Cb is higher than when it does not hold. In this case, the utility of the anaphor referring to Cb also tends to become high because of Preference 1b, so that the expected utility is high. Similarly, when the second condition holds, the reference probability of Cb and the utility of Cb are high, thus, the expected utility is also high. Furthermore, the first condition has stronger influence than the second, because the first one represents the cohesion between the previous and the current utterances, whereas the second merely predicts the cohesion between the current and the next utterances. Consequently, RETAIN has a larger expected utility than SMOOTH-SHIFT. Rule 2 of CT can thus be derived from the general principle of maximum expected utility, which is stated as follows:

Preference 2: The interlocutors prefer an interpretation of anaphora with higher expected utility.

This preference is a generalization of Rule 2.

We will verify the above preferences and provide evidence of the existence of the MG principle behind Rules 1 and 2 of CT in Section 5.

4. Definition and Measurement of Salience

Salience represents the likelihood of a discourse entity to be pronominalized or its degree of attention. In CT, it is not quantitatively defined despite that it plays critical roles in the theory. Therefore, we formulate salience in terms of a *reference probability* — a measurable quantity. The salience of an entity can be defined based on how probable the entity will also be referenced in the following utterances. In other words, if an entity seems to be referenced in the following discourse, the entity tends to draw attention and its salience can be considered to be high. This formulation resolves the issues of CT that were discussed in Section 2.

The salience of an entity e at the target utterance U_i is empirically defined as follows.

Definition of Salience: The salience of e at U_i is defined as the reference probability $\Pr(e, U_i)$, which is the probability of e being referenced in the next utterance U_{i+1}. Given a large amount of linguistic data, $\Pr(e, U_i)$ can be calculated as follows:

1. Find the latest reference to e in the previous discourse $[U_1, \ldots, U_i]$. Let it be w_e.
2. Compose the feature vector $feat(w_e, U_i)$ from w_e and $[U_1, \ldots, U_i]$. For example, the features we used in this study are listed in Table 3.
3. Extract samples (w_x, U_j) whose feature vectors $feat(w_x, U_j)$ equal $feat(w_e, U_i)$ from a large amount of linguistic data.
4. Using the extracted samples, calculate $\Pr(w_x, U_j)$, the probability that the referent of w_x is also referenced in U_{j+1} (in other words, calculate the relative frequency of samples that the referent of w_x is also referenced in U_{j+1}).
5. Take $\Pr(w_x, U_j)$ to be $\Pr(e, U_i)$.

Table 3: Features used in regression analysis of Pr(e, U_i)

(1)	*Dist*	*log* ((# utterances between U_i and the latest reference to e) + 1)
	Gram	grammatical function of the latest reference to e (*wa/ga/no/o/ni/mo/de/kara/to*)
	Chain	*log* ((# references to e in the previous context of U_i) + 1)
(2)	*Exp*	expression type of the latest reference to e (pronoun/non-pronoun)
	last_topic	whether the latest reference to e was the last topic (yes/no)
	last_sbj	whether the latest reference to e was the last subject (yes/no)
	p1	whether e was in the first person (yes/no)
	Pos	part of speech of the latest reference to e

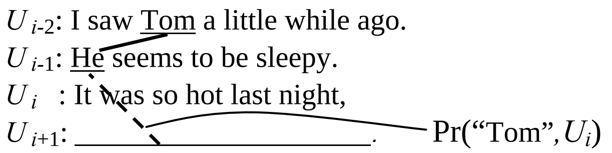
We used (1) for MLR, and (1) and (2) for SVR.

This definition is expressed by the following equation.

$$\text{(Salience of } e \text{ at } U_i) := \Pr(e, U_i) \approx \Pr(w_x, U_j) = \frac{\#\{(w_x, U_j)\,; C \wedge D\}}{\#\{(w_x, U_j); C\}}$$

$$\left[\begin{array}{l}\text{Condition } C\text{: } feat(w_x, U_j) = feat(w_e,\ U_i) \\ \text{Condition } D\text{: The referent of } w_x \text{ is also referenced in } U_{j+1}.\end{array}\right]$$

Hereafter, we explain the measurement of the salience of "Tom" at U_i for the following example.



$U_{i\text{-}2}$: I saw Tom a little while ago.

$U_{i\text{-}1}$: He seems to be sleepy.

U_i : It was so hot last night,

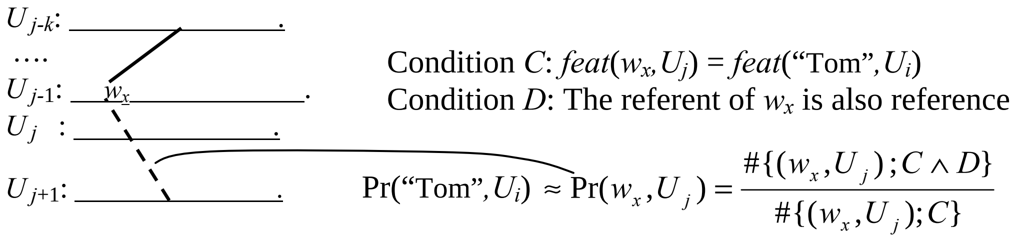
U_{i+1}: ________________. Pr("Tom", U_i)

In this example, the anaphor "he" refers to "Tom" and it appears in the last position among expressions referring to "Tom" in the previous discourse. We call it the *latest reference* to "Tom".

To simplify the explanation, if the following three features are used, *feat*("Tom", U_i) is defined as (*dist* = 2, *gram* =subject, *chain* = 2).

- *dist* : Utterances between U_i and the latest reference to e.
- *gram* : Grammatical function of the latest reference to e.
- *chain* : References to e in the previous context of U_i

We extract samples (w_x, U_j) that have the same feature vector as *feat*("Tom", U_i) from a corpus and calculate the reference probability from these samples.



$U_{j\text{-}k}$: ________________.

....

$U_{j\text{-}1}$: w_x________________.

U_j : ________________.

U_{j+1}: ________________.

Condition C: *feat*(w_x, U_j) = *feat*("Tom", U_i)

Condition D: The referent of w_x is also referenced in U_{j+1}.

$$\Pr(\text{"Tom"}, U_i) \approx \Pr(w_x, U_j) = \frac{\#\{(w_x, U_j)\,; C \wedge D\}}{\#\{(w_x, U_j); C\}}$$

Notice that interpolation and extrapolation are necessary because of data sparseness in the corpus. To this end, we used regression analysis for the measurements. We measured the reference probability with two regression algorithms: MLR (multiple logistic regression) and SVR (support vector regression). Table 3 lists the features for regression in this study.

5. Empirical Verification of Meaning-Game-based Generalization

We statistically verified Preference 1a, 1b, and 2 derived from MG. We used 1,356 articles taken from Japanese newspapers annotated with Global Document Annotation (GDA) (Hasida, 1998). These articles contained 63,562 utterances (predicate clauses). Table 4 shows the distribution of anaphora types in the corpus. To measure the reference probability, we extracted 1,073,781 samples of previously referenced entities for each utterance. Table 5 shows that there were 16,728 pairs of an utterance U_i and a previously referenced entity *e* that is also referenced in U_i (namely, that corpus contains 16,728 anaphors).

Table 4: Distribution of Japanese anaphora types

Anaphora Types	# Sample	Ratio
Zero Pronoun	5876	35.1%
Pronoun	843	5.0%
Noun Phrase with Demonstrative	1011	6.0%
Other Noun Phrase	8998	53.8%
Total	16728	100.0%

Table 5: Percentage of samples in which previously referenced *e* is also referenced in U_i

	# Sample (*e*, Ui)	Percentage
e is referenced in U_i	16,728	1.6%
e is not referenced in U_i	1,057,053	98.4%
Total	1,073,781	100.0%

We assumed that the utility of pronouns is greater than that of non-pronouns; i.e., the utility of pronouns equals 2, and that of non-pronouns equals 1. This assumption is equivalent to distinguishing between pronouns and non-pronouns in CT.

5.1 Measurement of Salience as Reference Probability

We measured reference probability, which is required in the verification of our MG-based generalization in Section 5.2. We used two regression algorithms for the measurement: MLR and SVR, which will be mentioned in Section 5.1.2 and 5.1.3, respectively. These regressions, especially MLR, require that their features must be numeric values. However, a grammatical function is not defined as numeric values. Therefore, we assigned numeric values to the

Table 6: Reference probabilities by only Japanese grammatical functions (by particles)

Particle (Grammatical Function)	# Sample	Referenced in U_{i+1}	Reference Probability
wa (topic)	35,329	1,908	0.0540
ga (subject)	38,450	1,107	0.0288
no (of)	88,695	1,755	0.0198
o (direct object)	50,217	898	0.0179
ni (indirect object)	46,058	569	0.0124
mo	8,710	105	0.0121
de	24,142	267	0.0111
kara	7,963	76	0.00954
to	19,383	129	0.00666
Other particles	512,006	8,027	0.0157
No particle	153,197	1,315	0.00858

grammatical functions as a preparation for the multiple regressions in Section 5.1.1.

After these regressions, we will also reconsider the conventional Japanese Cf-ranking based on the assigned values for grammatical functions in Section 5.1.4.

5.1.1 Assigning Numeric Values to Grammatical Functions

We measured the reference probabilities by using only grammatical functions for enabling the regression. This preparative measurement involves not regression but counting samples. Table 6 shows the reference probabilities calculated by from only the grammatical functions existing in the corpus. We used these values as *gram* in the multiple regression of the reference probability.

5.1.2 Measurement with MLR

MLR model is based on an assumption that the log odds of probability, $\log(P/(1-P))$, of some kind of event can be expressed as a linear expression of the explanatory variables. The regression function with three features in Table 3 is

$$\begin{aligned} Pr &= (1+\exp(-\lambda))^{-1} \\ &= (1+\exp(-(b_0 + b_1 dist + b_2 gram + b_3 chain)))^{-1}. \end{aligned}$$

It takes a huge amount of time to perform MLR on 1,073,781 samples. Thus, we made five regression models using 12,000 subsamples per model. We used statistical software called R (R Development Core Team, 2002) for the MLR analysis. Table 7 shows the parameters of the five regression models. We used average of probabilities predicted by the five models as the reference probability, which is represented as follows:

$$Pr = \frac{1}{5}\sum_{k=1}^{5}(1+\exp(-(b_{k,0} + b_{k,1} dist + b_{k,2} gram + b_{k,3} chain)))^{-1}$$

Table 7: Measured parameters of the five MLR models

k: Model No.	$b_{k,0}$ (const.)	$b_{k,1}$ (coeff. of *dist*)	$b_{k,2}$ (coeff. of *gram*)	$b_{k,3}$ (coeff. of *chain*)
1	-2.825	-0.7636	9.036	2.048
2	-3.055	-0.7067	10.47	2.270
3	-2.952	-0.7574	6.433	2.399
4	-3.288	-0.5911	9.170	2.129
5	-3.043	-0.6578	4.836	2.178

5.1.3 Measurement with SVR

We also made an SVR model to measure the reference probability with the eight features listed in Table 3. In MLR, the input values of the target variable are 0 or 1. However, in SVR, the input values must be smoothed as real numbers. We subsampled 60,000 samples, smoothed the input variables by using the k-NN method with k =100, and made an SVR model of a 2nd-degree polynomial kernel by using TinySVM (Kudo, 2002).

5.1.4 Reconsideration of Conventional Japanese Cf-ranking

Notice that the direct object is higher ranked than the indirect object in Table 6. This order is opposite from the conventional Japanese Cf-ranking order (Kameyama, 1998, Walker et al., 1994). Unfortunately, we have no way of telling which order is right from only this result.

To verify the order of the direct and indirect objects in Japanese, we need to consider another view point. To set the other ranking for this purpose, we calculate coefficients of linear regression of an anaphor's utility (in the next utterance) by using only grammatical functions (in the current utterance) in the corpus. These coefficients are assigned to each grammatical function for maximizing the correlation between them and utilities of noun phrases. In other words, they can be regarded as rigged values to maximize the predictive ability of Preferences 1a and 1b. Table 8 lists these coefficients and their ranking among the grammatical functions. This result also indicates that the direct object is higher ranked than the indirect object. Consequently, these empirical results disprove the order of direct and indirect objects in the conventional Japanese Cf-ranking.

Table 8: Coefficient for maximizing correlation to *Ut*

Particle (Grammatical Function)	Coefficient
wa (topic)	5.46
mo	5.37
ga (subject)	5.27
kara	5.14
o (object)	5.12
to	5.05
ni (indirect object)	5.05
no (of)	5.04
de	4.98

5.2 Empirical Verification of MG-based Generalization

We statistically verified Preferences 1a, 1b, and 2 using the reference probabilities obtained in Section 5.1.

5.2.1 Verification of Preference 1a and 1b

Firstly, we made pairs of anaphors in the same utterance for verifying Preference 1a. There were 914 pairs in utterances having multiple anaphors (Table 9). There were 360 pronoun and non-pronoun pairs. Thus, we calculated the percentage of samples that agreed with the prediction of

Table 9: Distribution of anaphor number in the same utterance

Anaphors in a same utterance	Utterances	Anaphors	Percentage	Pairs of anaphors in the same utterance
0	47,728	-	-	-
1	14,960	14,960	89.4%	-
2	854	1,708	10.2%	854
3	20	60	0.4%	60
Total	63,562	16,728	100.0%	914

Table 10: Verification of Preference 1a and 1b

		Measured	95% Confidence Interval
MLR	Preference 1a: Agreement Percentage (in 360 pairs of anaphor)	75.3% (271/360)	[70.5, 79.6]
	Preference 1b: Correlation Coefficient (in 16,728 samples)	+0.373	[0.360, 0.386]
SVR	Preference 1a: Agreement Percentage (in 360 pairs of anaphor)	74.4% (268/360)	[69.6, 78.9]
	Preference 1b: Correlation Coefficient (in 16,728 samples)	+0.386	[0.373, 0.399]

Preference 1a from these 360 pairs. We assumed that the percentage was binomially distributed and calculated the 95% confidence interval of the percentage. Table 10 shows that at least 70% of the samples agreed with Preference 1a.

Secondly, to verify Preference 1b, we measured Pearson's correlation coefficient between reference probability and utility. We assumed that the correlation coefficient was t-distributed and calculated the 95% confidence interval. Table 10 shows that the correlation coefficient of Preference 1b was at least +0.36 in both MLR and SVR. These results show that Preferences 1a and 1b were verified with statistical significance.

5.2.2 Verification of Preference 2

Table 11 shows the distribution of transition states in the corpus (decision on centers based on reference probabilities estimated by MLR as salience values). We see that the frequency of RETAIN is low despite its high preference rank. This tendency of the paucity of RETAIN has also been observed by Iida (1997) and Yamura-Takei et al. (2000). We cannot assume that the preference order matches the frequency order because the four transition states can not always be selected for every utterance.

Table 12 shows the averages and variances of the expected utility for each transition state. The order of the data in the table conforms to the order of Rule 2 of CT. We tested this order with multiple comparison tests. The results of the Kruskal-Wallis test were χ^2 = 1780.7, *df* = 3, and *P* $< 2.2\times 10^{-16}$. This means that the averages of the four states differed significantly. Table 13 shows the result of the Wilcoxon's rank sum test using the method of Holm, which demonstrates that the order is statistically significant. Additionally, the correlation coefficient between the transition states and the averages of expected utility in Table 12 was +0.520 when we assigned the following

Table 11: Distribution of transition states

	CONTINUE	RETAIN	SMOOTH-SHIFT	ROUGH-SHIFT
Zero Pronoun	56.0%(1315/2347)	1.7%(41/2347)	38.3%(898/2347)	4.0%(93/2347)
Pronoun	43.6%(102/234)	2.1%(5/234)	50.9%(8/234)	3.4%(8/234)
Total of Pronoun	54.9%(1417/2581)	1.8%(46/2581)	39.4%(1017/2581)	3.9%(101/2581)
Noun Phrase with Dem.	20.9%(56/268)	3.0%(8/268)	64.2%(172/268)	11.9%(32/268)
Other Noun Phrase	20.0%(522/2611)	1.8%(48/2611)	67.4%(1761/2611)	10.7%(280/2611)
Total of Non-Pronoun	20.1%(578/2879)	1.9%(56/2879)	67.1%(1933/2879)	10.8%(312/2879)
Total of All	36.5%(1995/5460)	1.9%(102/5460)	54.0%(2950/5460)	7.6%(413/5460)

Table 12: Averages of expected utilities in each transition states

Transition State	#Sample	Ave. of Expected Utility (Their Variance)
CONTINUE	1,995	0.874 (0.361)
RETAIN	102	0.473 (0.242)
SMOOTH-SHIFT	2,950	0.287 (0.175)
ROUGH-SHIFT	413	0.109 (0.0336)

Table 13: Wilcoxon's rank sum test

Pair of Transition States for Comparison	Significance Probability
CONTINUE : RETAIN	5.89×10^{-13}
CONTINUE : SMOOTH-SHIFT	$< 2.2 \times 10^{-16}$
CONTINUE : ROUGH-SHIFT	$< 2.2 \times 10^{-16}$
RETAIN : SMOOTH-SHIFT	1.64×10^{-6}
RETAIN : ROUGH-SHIFT	$< 2.2 \times 10^{-16}$
SMOOTH-SHIFT : ROUGH-SHIFT	$< 2.2 \times 10^{-16}$

values: CONITNUE=4, RETAIN=3, SMOOTH-SHIFT=2, ROUGH-SHIFT=1. This means that expected utility correlates with topic continuity.

These results provide statistical evidence of a principle of the Meaning Game behind Rules 1 and 2 in Centering Theory.

6. Discussion

6.1 Effect of Quantitative Definition of Salience as the Reference Probability

We resolved the problems regarding salience in Section 2 by formulating it as a reference probability. That is,

- In our formulation, salience becomes a measurable quantity and becomes statistically verifiable based on large linguistic data.
- By adopting regression algorithms that can handle multiple explanatory variables, the model can more easily integrate features that influence salience than heuristic methods of previous works.
- By taking into account the distance between the current utterance and the latest references to entities, the model can handle not only entities referenced in the previous utterance but also all entities in the previous discourse.

6.2 Samples Disagreeing with Preference 1a

In Section 5.1.1, we confirmed that 75.3% of samples favored Preference 1a with MLR and that 74.4% samples favored it with SVR. This means, however, that about 25% of samples did not favor Preference 1a. In checking the results, we found that semantic features (e.g., selectional restriction of predicate, semantic category of anaphor, and so on.) accounted for this difference. We consider that if these features were imported to the multiple regression of the reference probability, Preferences 1a and 1b would become even stronger.

6.3 Strictness of Verification of Preference 2

In Section 5.2.2, we verified the correspondence between the order of averages expected utility for each transition state and the order of Rule 2 of CT. This means that Rule 2 of CT can be derived from MG. This verification is, however, not always strict. We should verify the order of the solutions of each case, not the order of the averages. We leave this issue as a future work.

7. Conclusion

CT has two limitations despite it being a standard theory about discourses. It lacks a principle behind discourse phenomena and a quantitative definition of salience. We have quantitatively formulated salience as a reference probability from the standpoint that the principle underlying discourse can be attributed to game theory. Furthermore, we formulated two preferences as the MG-based generalization of CT and statistically verified these preferences in a Japanese corpus.

In our verification of Preference 1a and 1b, we claimed that there is a positive correlation between the utility of an anaphor and the reference probability of its referent. In connection with this, the conventional Japanese Cf-ranking was empirically disproved. Preferences 1a and 1b derived from MG cover more general cases than Rule 1 in CT does. In our verification of Preference 2, we estimated the average expected utility for the four transition states and presented evidence that the order among the averages corresponds to that of Rule 2 in CT. These empirical results indicate that a principle informed by MG is behind both rules in CT.

We hence conclude that we have statistically proved our MG-based generalization. It is a more quantitative and principled model than conventional CT.

8. Acknowledgements

The authors are grateful to members of the now-defunct Cyber Assist Research Center and the people who annotated the GDA tags to the corpus used in our study.

9. References

Grosz, B.J., Joshi, A.K., and Weinstein, S. 1995. Centering: A framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), pp. 203-225.

Hasida, K., Nagao, K., and Miyata, T. 1995. A Game-Theoretic Account of Collaboration in Communication. *Proceedings of the First International Conference on Multi-Agent Systems,* pp. 140-147.

Hasida, K. 1996. Issues in Communication Game. *Proceedings of the 16th Conference on Computational Linguistics,* pp. 531-536.

Hasida, K. 1998. Global Document Annotation (GDA). http://i-content.org/GDA/.

Iida, M.. 1997. Discourse Coherence and Shifting Centers in Japanese Texts. In M. Walker, A. Joshi, and E. Prince, eds., *Centering Theory in Discourse*, pp. 161-180.

Kameyama, M. 1997. Intrasentential Centering: A Case Study. In M. Walker, A. Joshi, and E. Prince, eds.*, Centering in Discourse,* pp. 89-112.

Kudo, T. 2002. TinySVM: Support Vector Machines. http://chasen.org/~taku/software/TinySVM/.

Nariyama, S. 2001. Multiple Argument Ellipses Resolution in Japanese. *Proceedings of Machine Translation Summit VIII*, pp. 241-245.

Poesio, M., Stevenson, R., Di Eugenio, B., and Hitzeman, J. 2004. Centering: A Parametric Theory and Its Instantiations. *Computational Linguistics*, 30(3), pp. 309-363.

R Development Core Team. 2002. The R Project for Statistical Computing. http://www.r-project.org/.

Strube, M. and Hahn, U. 1999. Functional Centering: Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3), pp. 309-344.

Yamura-Takei, M., Takada, M., Aizawa, T. 2000. The Role of Global Topic in Japanese Zero Anaphora Resolution (in Japanese). *Technical Report, of IPSJ*, 135(10), pp. 71-78.

Walker, M.A., Iida, M., and Cote, S. 1994. Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2), pp. 193-232.