Meaning-Game-based Centering Model with Statistical Definition of Utility of Referential Expression and Its Verification Using Japanese and English Corpora

Shun Shiramatsu*, Kazunori Komatani*, Kôiti Hasida†, Tetsuya Ogata*, Hiroshi G. Okuno*

*Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{siramatu, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp
† ITRI, National Institute of Advanced Industrial Science and Technology (AIST)
Akihabara Dai-Building, Soto-Kanda 1-18-13, Chiyoda-ku, Tokyo 101-0021, Japan
hasida.k@aist.go.jp

Abstract

This paper presents a quantitative modeling of referential coherence by which conversation systems measure the smoothness of discourse. Investigations of the corpora show that referential coherence depends on languages or genres of discourse. Our goal is to establish a quantitative model that can be statistically adapted to various languages. Centering theory explains referential coherence by using heuristic rules. Since these heuristics should be invented manually for a particular language, we need a quantitative/statistical model that can be obtained from a corpus. The meaning-game-based centering model (MGCM) (Shiramatsu et al., 2005) quantitatively reformulates centering theory by exploiting quantitative aspects of game theory. It quantifies referential coherence by using the two parameters related to salience and pronominalization: "reference probability" and "perceptual utility". However, MGCM still has two problems. The first is that perceptual utility cannot be statistically adapted to various languages. The second is that MGCM has only been verified in Japanese. We have enhanced the model by statistically defining perceptual utility. Specifically, we defined it by using occurrence frequency of the referential expression in a corpus. Experimental results using English and Japanese corpora showed that our statistical definitions enabled the parameters to be adapted to both corpora. Furthermore, the statistical tests of the enhanced MGCM showed its validity in both corpora.

1. Introduction

Quantification of referential coherence is important for conversation systems to be able to automatically measure the smoothness of generated discourse (e.g., when the system selects a coherent utterance from a number of candidate utterances). Referential coherence can be measured on the basis of discourse salience and pronominalization. Referential coherence depends on languages or genres of discourse, according to investigations of corpora. Our goal is to formulate a quantitative model of referential coherence with parameters related to salience and pronominalization. We assume that pronominalization can be quantitatively defined as the amount of perceptual load reduction in the interlocutors' cognitive process. It is desirable that the model be applicable to various languages. To adapt the model to various language corpora, the parameters of the model should be statistically defined.

Centering theory (CT) (Grosz et al., 1995) is a theory of discourse salience, anaphora, and referential coherence based on heuristic rules (e.g., Cf-ranking and transition ranking). It is useful for estimating referential coherence between utterance units and for selecting appropriate referential expressions. It does not, however, give a quantitative model.

The resolution of anaphora procedure (RAP) (Lappin and Leass, 1994), another approach similar to CT, includes salience weightings based on Alshawi (1987)'s salience factors. Although its salience weightings is apparently quantitative, it is defined as a rule-based weighting with a priori heuristics, without corpus-based statistics.

The meaning-game-based centering model (MGCM) (Shiramatsu et al., 2005) is a quantitative reformulation of CT, which is derived from game theory. It quantifies referential coherence with a principle of expected utility, which is game-theoretical hypothesis about the relationship between discourse salience and pronominalization.

Because this principle is not based on the properties of a specific language, the model can be applied to various languages. Furthermore, such a quantitative approach can obtain parameters optimized to a target language by statistically analyzing a corpus in that language.

The modeling of MGCM, however, is not fully statistical. Specifically, the *perceptual utility* of referential expressions, which represents reduction of the perceptual load, is manually designed. Such a manual approach requires specialized skills to adapt the model to each language. Moreover, such manual adaptation is difficult to justify its accuracy. Due to this imperfection of the methodology of adaptation, the effectiveness of MGCM has so far been verified only with a Japanese corpus. For verification in other languages, it requires corpus-based statistical design in order to enable easier and more accurate parameter fitting.

In this paper, we statistically define perceptual utility. The definition is based on the assumption that higher-utility words should be costless and frequently used (i.e., more familiar) in the target corpus. Furthermore, we verify the model using both Japanese and English corpora. We quantitatively discuss the difference in pronominalization between Japanese and English corpora.

Table 1: Transition types

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

Here, $\mathrm{Cp}(U_i)$, the preferred center of utterance U_i, is the highest ranked element of $\mathrm{Cf}(U_i)$.

2. Problems with Conventional Studies

This section describes the issues of conventional CT and MGCM.

2.1. Centering Theory

CT handles a discourse as a sequence of utterance units $[U_1, U_2, \cdots, U_n]$. CT explains the referential expression between U_{i+1} and U_i by using heuristics: a salience ranking of grammatical roles (Cf-ranking) (Walker et al., 1994), three constraints about "center", and two rules about referential coherence (Poesio et al., 2004).

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

Cb(U_i): The backword-looking center of U_i.
Cf(U_i): The forward-looking centers of U_i.

Constraint 1: All utterances of a segment except for the 1st have at most one Cb.
Constraint 2: Every element of $\mathrm{Cf}(U_i)$ must be *realized* in U_i.
Constraint 3: $\mathrm{Cb}(U_i)$ is the highest ranked element of $\mathrm{Cf}(U_{i-1})$ that is realized in U_i.

Rule 1 (Pronominalization rule): If any $\mathrm{Cf}(U_i)$ is pronominalized, $\mathrm{Cb}(U_i)$ is also pronominalized.
Rule 2 (Transition rule): Transition types are ordered: CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT (See also Table 1).

Different researchers proposed different versions of the above heuristics. The variations of them, however, do not clearly specify a principle behind referential coherence represented with Rules 1 and 2. We fear that variations of CT will disorderly grow without principled background. We need a simple modeling which explains principled mechanism behind the phenomena.

Moreover, the Cf-ranking requires specialized skills for adapting it to each language. For example, the ordering of objects and indirect objects in Japanese is difficult to justify. We need a methodology for automatic adaptation.

2.2. Meaning-Game-based Centering Model

The meaning game (MG) (Hasida, 1996) is a game-theoretical model of intentional communication. MGCM (Shiramatsu et al., 2005) is a quantitative reformulation of CT based on the MG framework. It quantifies referential coherence with the following two parameters.

- **Reference probability (Pr):** Probability of a target entity being referenced in a following utterance unit. It represents discourse salience, i.e., degree of attention to the target discourse entity.
- **Perceptual utility (Ut):** Reduction of perceptual cost when interlocutors transmit a target referential expression. For example, ellipses and pronouns have higher perceptual utilities because they are perceptually simple and costless.

Table 2: Correspondence of CT and MGCM

	CT (Non-quantitative)	MGCM (Quantitative)
Discourse salience	Cf-ranking (Subject>Object> $\cdots$)	**Reference probability** $\Pr(e\|U_1, \cdots, U_i)$
Load reduction	Pronominalization (Pronoun / Non-pronoun)	**Perceptual utility** $\mathrm{Ut}(w)$
Referential coherence	Transition ranking (CONTINUE >RETAIN >SMOOTH-SHIFT >ROUGH-SHIFT)	**Expected utility** $\mathrm{EU}(U_{i+1}) = \sum_{w \text{ refers to } e \text{ in } U_{i+1}} \Pr(e\|U_1, \cdots, U_i)\mathrm{Ut}(w)$

$U_1, \cdots, U_i$: Preceding discourse
U_{i+1}: Following utterance unit
e: Entity referenced in $U_1, \cdots, U_i$
w: Referential expression in U_{i+1}

Table 3: Features of reference pattern of e used in calculating $\Pr(e|U_1, \cdots, U_i)$

dist	log ((# utterances between U_i and the latest reference to e in $[U_1, \cdots, U_i]$) + 1)
gram	grammatical role of the latest reference to e in $[U_1, \cdots, U_i]$
chain	log ((# references to e in $[U_1, \cdots, U_i]$) + 1)

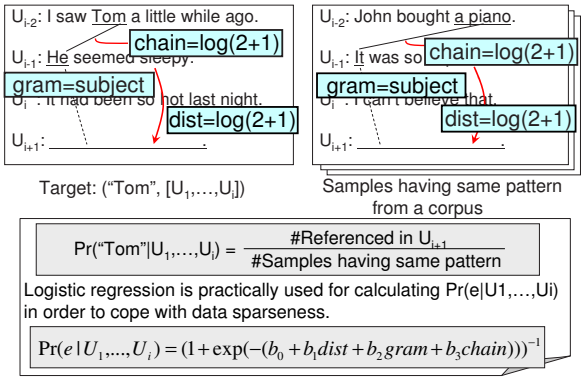

Figure 1: Calculation of $\Pr(e|U_1, \cdots, U_i)$

Table 2 presents the correspondence between MGCM and CT. Here, e represents a target entity referenced in a preceding discourse, $[U_1, \cdots, U_i]$. w represents a referential expression referring to e in the following utterance unit, U_{i+1}.

The reference probability of e, $\Pr(e|U_1, \cdots, U_i)$, represents the conditional probability of e being referenced in U_{i+1}, given the reference pattern of e in $[U_1, \cdots, U_i]$. Table 3 shows the features we used as the reference pattern. Figure 1 outlines the calculation of the Pr value by using a corpus. The calculation uses logistic regression analysis in order to cope with data sparseness. Pr represents the degree of salience of e at the moment of U_i.

The perceptual utility of w, Ut(w), represents the perceptual load reduction of using the referential expression w. For example, ellipsis and pronouns have higher utility because their perceptual cost is lower than other noun phrases. Ut(w) is used for generalizing Rule 1 of CT, which is a rule about pronominalization.

The expected utility of U_{i+1}, EU(U_{i+1}), represents the expected load reduction when interlocutors output or predict U_{i+1}. For example, EU(U_{i+1}) is high when a costless pronoun in U_{i+1} refers to a salient entity. It represents the degree of referential coherence between U_{i+1} and $[U_1, \cdots, U_i]$.

The principle of MGCM is that referential coherence is achieved by reducing the communicating load for interlocutors, which is represented as the expected utility derived from game theory. This principle is possibly common to various languages; hence MGCM should be applied to the various languages in order to find out if this is the case. To apply it to various languages, two parameters, Pr and Ut, should be statistically obtained from a corpus of each language.

Shiramatsu et al. (2005) statistically defined reference probability, Pr, in order to enable corpus-based fitting of Pr. Perceptual utility, Ut, however, was not statistically defined. It was naively assumed that a pronoun had higher Ut than a non-pronoun. A corpus-based statistical definition of Ut is needed to apply MGCM to other language corpora.

In fact, the effectiveness of the model has so far been verified only with a Japanese corpus.

3. Statistical Definition of Perceptual Utility

Perceptual utility represents interlocutors' perceptual load reduction in transmitting (speaking, writing, hearing, or reading) a lexical symbol of a target referential expression. It does not contain the cost of semantic understanding (e.g., anaphor resolution). Shiramatsu et al. (2005) naively assumed that pronouns had higher perceptual utilities than non-pronouns.

We assume that perceptual simplicity can be calculated using frequency of referential expressions for the following reasons: Frequent expressions cost less than rare ones because interlocutors are habituated to frequent ones. Costless expressions are more frequently used than costly ones.

We define the perceptual cost of a target referential expression w as follows:

(Perceptual cost of w) $= \mathrm{I}(w) = -\log p(w)$.

Here, $p(w)$ represents the probability of w appearing as an anaphoric expression in a corpus. Given this definition, the perceptual cost of w is calculated with the following equation:

$$\begin{aligned}\mathrm{I}(w) &= -\log p(w) \\ &= -\log \frac{\#w \text{ as anaphoric expressions}}{\#\text{ all utterance units}}\ [nat].\end{aligned}$$

The perceptual utility is the reverse of the perceptual cost. We define the perceptual utility of a target referential expression w as follows:

(Perceptual utility of w) $= \mathrm{Ut}(w)$
$= -$(perceptual cost of w) + (basic level)
$= -I(w) + Ut_0$.

The basic level Ut_0 is empirically defined to ensure that $\mathrm{Ut}(w) > 0$.

We propose to replace the definition of perceptual utility in Shiramatsu et al. (2005) with the new one. We call this enhanced version "enhanced MGCM" hereafter.

4. Empirical Verification using Large Japanese and English Corpora

To verify enhanced MGCM on different languages, we used the Wall Street Journal (hereafter, *WSJ*) as an English corpus and the Mainichi-Shinbun (hereafter, *Mainichi*) as a Japanese corpus. *WSJ* contains 2,412 articles, 135,278 predicate clauses, and 95,677 anaphors. *Mainichi* contains 1,356 articles, 63,562 predicate clauses, and 16,728 anaphors. The verification requires linguistic annotations which specify structure of morpheme, dependency, and anaphora. Both corpora are manually annotated according to Global Document Annotation (GDA) (Hasida, 1998). The following examples illustrate the GDA tags specifying the anaphora structures.

```
<su syn="b">
  <namep id="Foo">The foo model </namep>
  should be adaptable to
  <np id="Data">various data</np>.</su>
<su syn="b">
  <adp>However</adp>, <np eq="Foo">it </np>
  lacks the methodology for
  <np obj="Foo" gol="Data">adapting</np>.</su>
```

Here, the attribute `id` represents antecedent. The attribute `eq` represents anaphor except for ellipsis. The relational attributes `obj` and `gol` represent ellipsis.

4.1. Verification of Definition of Discourse Salience

The MGCM statistically defined discourse salience as a reference probability Pr($e|U_1, \cdots, U_i$). Calculation of Pr (See also Figure 1) required the following two preparations. Firstly, we needed to assign the *gram* value to each grammatical role. We assigned Pr average of each grammatical role, Pr(*gram*), which was calculated by counting samples in *Mainichi* and *WSJ* corpora (Tables 4 and 5). Secondly, we needed to obtain the regression weights b_i from the corpora. We obtained b_i by logistic regression analysis (Table 6). We used 12,000 subsamples from corpora per one trial of the logistic regression analysis.

Tables 4 and 5 show the consistency between Pr(*gram*) ranking and the conventional Cf-ranking. The Pr order in *Mainichi* among "Topic," "Subject," and "Object" was consistent with the conventional Japanese Cf-ranking. The Pr order in *WSJ* among "Subject," "Object," and "Complement" was also consistent with the conventional English Cf-ranking. These consistencies indicate the validity of the Pr definition in the both corpora.

Tables 4 and 5 also illustrate the difference between *Mainichi* and *WSJ*. Although they were similar in the Pr order between "Subject" and "Object", they were different in the distributions of Pr value. This indicates that Pr was

Table 4: Reference probability for each grammatical role (*Mainichi*, Japanese corpus)

Type of *gram*	# Samples	# Referenced	Pr(*gram*)
Topic (*wa*)	35,329	1,908	5.40×10^{-2}
Subject (*ga*)	38,450	1,107	2.88×10^{-2}
(*no*)	88,695	1,755	1.98×10^{-2}
Object (*o*)	50,217	898	1.79×10^{-2}
Indirect object (*ni*)	46,058	569	1.24×10^{-2}
(*mo*)	8,710	105	1.21×10^{-2}
(*de*)	24,142	267	1.11×10^{-2}
(*kara*)	7,963	76	9.54×10^{-3}
(*to*)	19,383	129	6.66×10^{-3}

Table 5: Reference probability for each grammatical role (*WSJ*, English corpus)

Type of *gram*	# Samples	# Referenced	Pr(*gram*)
Subject	76,147	16,441	2.16×10^{-1}
(*by*)	5,045	618	1.22×10^{-1}
Indirect object	1,569	184	1.17×10^{-1}
(*with*)	4,272	446	1.04×10^{-1}
(*of*)	23,798	2,145	9.01×10^{-2}
(*from*)	4,005	350	8.74×10^{-2}
Object	42,578	3,703	8.70×10^{-2}
(*to*)	8,449	661	7.82×10^{-2}
(*for*)	7,759	601	7.75×10^{-2}
(*on*)	5,140	229	5.82×10^{-2}
(*at*)	4,043	233	5.76×10^{-2}
Complement	7,102	371	5.22×10^{-2}

Table 6: Regression weights in logistic regression for Pr

Corpus	b_0(Const.)	b_1(*dist*)	b_2(*gram*)	b_3(*chain*)
Mainichi	-2.825	-0.7636	9.036	2.048
WSJ	-2.405	-1.411	8.788	3.519

fitted to each corpus. Additionally, the difference is due to language difference in the types of grammatical roles. This also indicates the necessity of corpus-based parameter fittings for each language.

Table 6 shows the consistency between the regression weights obtained from the corpora and linguistic heuristics. The negative values of the weight of *dist*, b_1, are consistent with the heuristics that the recently referenced entities are more salient than the earlier ones. The positive values of the weight of *gram*, b_2, are consistent because Pr(*gram*) represents salience of each grammatical role. The positive values of the weight of *chain*, b_3, are consistent with the heuristics that the frequently referenced entities are more salient than the rare ones. These consistencies in both corpora indicate that the adaptation to the corpora is successful by logistic regression analysis. These also indicate the validity of the Pr definition in both corpora.

4.2. Verification of Definition of Perceptual Cost

The enhanced MGCM statistically defines the perceptual cost as I(w), the amount of self-information. We therefore measured I(w) for each referential expression w. Tables 7 and 8 show the rankings of referential expressions in order of perceptual cost.

An ellipsis (zero pronoun or empty category) has the lowest cost in both corpora. Pronouns (in colored rows in the tables) tend to have lower cost than non-pronouns in both corpora. The lower three rows in the tables list the average costs for ellipsis, pronoun, and other nouns. In both corpora, the rankings of the categories were consistent with the heuristics as follows:

Ellipsis $<$ Pronoun $<$ Other nouns

These results justify the validity of our definition in both *Mainichi* and *WSJ*.

Table 7: Perceptual cost of each referential expression (*Mainichi*, Japanese corpus)

Referential expression w	Appearance probability $p(w)$	Perceptual cost $I(w)$ [*nat*]
(Zero pronoun)	2.940×10^{-1}	1.224
watashi (I)	5.129×10^{-3}	5.273
sono (that)	3.965×10^{-3}	5.530
kore (this)	2.973×10^{-3}	5.818
kono (this)	1.888×10^{-3}	6.272
Nihon (Japan)	1.809×10^{-3}	6.315
mono (thing)	1.809×10^{-3}	6.315
⋮	⋮	⋮

Type of w	Average of $p(w)$	Perceptual cost
Zero pronoun	2.940×10^{-1}	1.224
Pronoun	2.403×10^{-3}	6.031
Other noun	2.271×10^{-4}	8.390

Table 8: Perceptual cost of each referential expression (*WSJ*, English corpus)

Referential expression w	Appearance probability $p(w)$	Perceptual cost $I(w)$ [*nat*]
(Empty category)	2.547×10^{-1}	1.368
it	4.232×10^{-2}	3.162
he	3.049×10^{-2}	3.490
they	1.850×10^{-2}	3.990
company	1.652×10^{-2}	4.103
we	1.112×10^{-2}	4.499
I	1.020×10^{-2}	4.585
U.S.	8.342×10^{-3}	4.786
⋮	⋮	⋮

Type of w	Average of $p(w)$	Perceptual cost
Empty category	2.457×10^{-1}	1.368
Pronoun	3.257×10^{-2}	3.836
Other noun	1.317×10^{-3}	6.632

Tables 7 and 8 also show the difference between *Mainichi* and *WSJ*. Although they were similar, the distributions were different due to language difference in the types of referential expressions. This indicates that Ut was fitted to each corpus. It also indicates the necessity of corpus-based parameter fitting to each language.

4.3. Verification of Preferences in Meaning-Game-based Centering Model

MGCM contains Preferences 1a, 1b, and 2, which are general formulations of Rules 1 and 2 of CT (Shiramatsu et al., 2005).

- **Preference 1a:** w_1 refers to e_1 and w_2 refers to e_2 when Pr($e_1|U_1,\cdots,U_i$) $>$ Pr($e_2|U_1,\cdots,U_i$) and Ut(w_1) $>$ Ut(w_2), given that (w_1, w_2) is a pair of anaphor in U_{i+1} (Figure 2).
- **Preference 1b:** There is a positive correlation between Pr($e|U_1,\cdots,U_i$) and Ut(w), given that w refers to e in U_{i+1} (Figure 2).

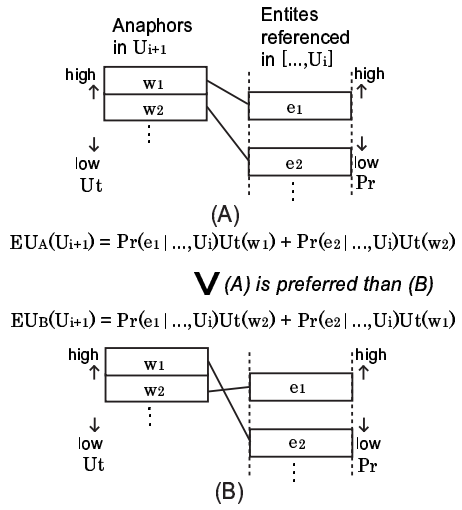

Figure 2: EU(U_{i+1}) increases through positive correlation between Pr and Ut (Preferences 1a and 1b).

- **Preference 2:** The higher EU(U_{i+1}) is preferred.

Preference 2 is the principle of MGCM derived from game theory. It is a generalization of Rule 2 of CT (i.e., transition rule). Preference 1a is derived from Preference 2 (See also Figure 2). Preference 1b is derived from Preference 1a. Preferences 1a and 1b are generalizations of Rule 1 of CT (i.e., pronominalization rule).

As preparation of verification, we had to determine the value of Ut_0, the basic level of Ut. We empirically determined that $Ut_0 = 12$[*nat*]. The grounds for this setting are described in the discussion section.

Verification of Preference 1a: We measured the ratio of samples which comply with Preference 1a (i.e., (A) in Figure 2) in order to verify Preference 1a. The ratio is influenced by differences in Pr and Ut between w_1 and w_2, i.e., $\Delta\mathrm{Pr} = \log\mathrm{Pr}(e_1) - \log\mathrm{Pr}(e_2)$ and $\Delta\mathrm{Ut} = \mathrm{Ut}(w_1) - \mathrm{Ut}(w_2)$. The greater the differences, the larger the ratio of the compliant samples.

We varied a threshold of $\Delta\mathrm{Pr}$ and $\Delta\mathrm{Ut}$ to investigate the influence of $\Delta\mathrm{Pr}$ and $\Delta\mathrm{Ut}$. We applied Preference 1a to samples only in range that $\Delta\mathrm{Pr}$ and $\Delta\mathrm{Ut}$ were greater than a certain threshold. Here, we took the ratio of in-range compliant samples to all compliant samples as the recall of Preference 1a. We took the ratio of in-range compliant samples to all in-range samples as the precision of Preference 1a. Figure 3 shows the recall-precision curve by the threshold varying. When the recall was 100%, the precision was greater than 60% on both corpora. When the recall was 60%, the precision was approximately 70% on both. These results indicate the validity of Preference 1a because the precisions are always greater than 60% over the whole range.

Verification of Preference 1b: We verified Preference 1b, i.e., the positive correlation between Pr and Ut. The correlation coefficient in *Mainichi* is greater than $+0.356$ at the 2.5% significance level. That in *WSJ* is greater than $+0.217$ at the 2.5% significance level. Therefore, Preference 1b was statistically significant in both corpora because the positive correlation was statistically significant.

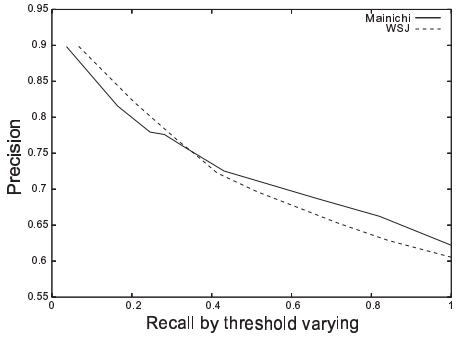

Figure 3: Recall-precision curves of Preference 1a

Table 9: Average of EU(U_{i+1}) for each transition type (*Mainichi*, Japanese corpus)

Transition type	Samples	Average of EU(U_{i+1})
CONTINUE	1,783	10.374
RETAIN	84	7.913
SMOOTH-SHIFT	2,704	2.624
ROUGH-SHIFT	194	1.428

Table 10: Average of EU(U_{i+1}) for each transition type (*WSJ*, English corpus)

Transition type	Samples	Average of EU(U_{i+1})
CONTINUE	13,384	5.439
RETAIN	2,314	3.295
SMOOTH-SHIFT	18,904	2.664
ROUGH-SHIFT	5,628	1.031

Verification of Preference 2: We verified the consistency between the expected utility, EU(U_{i+1}), of MGCM and the transition ranking, i.e., Rule 2 of CT. As preparation, we determined the transition types of samples in corpora by using Pr values as a substitute for the Cf-rankings.

Tables 9 and 10 show the consistency of ranking by average expected utility of MGCM with the transition ranking of CT. Wilcoxon's rank sum test was statistically significant on both corpora at the 99% confidence level. Therefore, Preference 2 was statistically significant in *Mainichi* and *WSJ*.

Furthermore, we investigated the correlation coefficient between the transition ranking and EU(U_{i+1}). As preparation, we assigned the ranked values according to transition ranking of CT: CONTINUE: 4, RETAIN: 3, SMOOTH-SHIFT: 2, and ROUGH-SHIFT: 1. As the result, the correlation coefficient in *Mainichi* was equal to +0.585. That in *WSJ* was equal to +0.407. These results also provide statistical evidence for the validity of EU(U_{i+1}) as a scale of the referential coherence between U_{i+1} and $[U_1, \cdots, U_i]$.

5. Discussions

Here, we quantitatively compare *Mainichi* and *WSJ*. Additionally, we describe the grounds for the Ut_0 setting.

5.1. Quantitative Comparison of *Mainichi* and *WSJ*

Here, we quantitatively compare *Mainichi* and *WSJ* from the viewpoint of Preference 1b. Although the correlation coefficients were significantly positive in the both corpora, the coefficient in *WSJ* was less than that in *Mainichi*. Figure 5 indicates its reason. It represents the correlation between

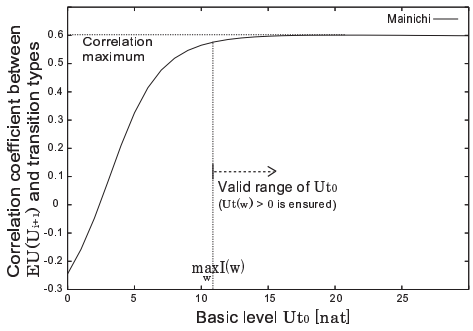

Figure 4: Valid range of Ut_0 setting in terms of consistency with Rule 2 of CT

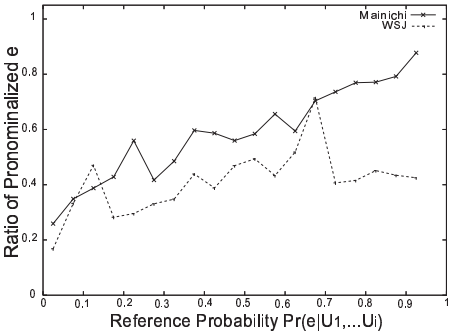

Figure 5: Ratios of pronominalized entities

the Pr value and the ratio of pronominalized entities, i.e., high-Ut entities. In the range of $\Pr < 0.75$, the ratio of pronominalized entities increased though the increase of Pr in both corpora. In the range of $\Pr > 0.75$ (i.e., the range of the salient entities), the correlations were, however, different between the corpora. In *Mainichi*, the pronominalization ratio smoothly increased through the increase of Pr in this range. In *WSJ*, the pronominalization ratio did not increase in this range, as contrasted with *Mainichi*.

We investigated this difference between *Mainichi* and *WSJ* in the range of the salient entities. In *Mainichi*, only 17.6% samples were not pronominalized in that range. In *WSJ*, 55.3% (11,367) samples were not pronominalized in that range as contrasted with *Mainichi*. By closely investigating, 41.7% (4,735) samples in the non-pronominalized samples were referenced by the proper nouns in *WSJ*.

In Japanese, the salient entities are frequently referenced by the zero pronouns. In English, especially in the newspaper articles, the salient entities are comparatively referenced by the contracted names (e.g., "Dr. Talcott" instead of "he") or the definite noun phrases. In this respect, there is still room for improvement in the definition of perceptual utility.

5.2. Grounds for the Ut_0 setting

We investigated the influence of Ut_0, the basic level of Ut. Figure 4 shows the correlation coefficient between EU(U_{i+1}) and the transition ranking in *Mainichi*.

This correlation relates to the consistency of Preference 2 with Rule 2 of CT. [1] The correlation was saturated when $Ut_0 \geq \max \mathrm{I}(w)$ (i.e., $Ut(w) \geq 0$ is ensured). However, it dramatically decreased when $Ut_0 < \max \mathrm{I}(w)$ (i.e., $Ut(w) \geq 0$ is not ensured). This result shows that Preference 2 is valid as long as $Ut(w) \geq 0$ is ensured. In *WSJ*, $\max \mathrm{I}(w) = 11.82$. In *Mainichi*, $\max \mathrm{I}(w) = 11.06$ [*nat*]. Thus, we set $Ut_0 = 12$ [*nat*] on both corpora.

[1] Rule 1 of CT does not depend on Ut_0 varying.

6. Conclusion

We enhanced the design of MGCM in order to establish a quantitative model adaptable to different language corpora. Two parameters, reference probability and perceptual utility, should be statistically adapted to various language corpora. We statistically defined perceptual utility of referential expressions as the load reduction by using the occurrence frequency in the corpus. In this way, we made the two parameters adaptable to corpora of various language.

Although the two parameters were distributed differently between Japanese and English corpora (Tables 4, 5, 7, and 8), our definitions of them were valid on both corpora. This indicates that optimal parameters can be obtained from a corpus of the target language.

The preferences of MGCM derived from the principle of expected utility were also valid on both corpora. Preferences 1a and 1b related to pronominalization, which are represented as positive correlations between reference probability and perceptual utility, were statistically significant. Preference 2 related to transition, which is represented as the principle of expected utility, was also statistically significant. These results indicate that MGCM and its principle are cross-linguistically valid in Japanese and English. They also indicate that the expected utility is a valid scale of referential coherence in Japanese and English.

Therefore, we confirmed that the enhanced MGCM was adaptable to Japanese and English corpora. Consequently, we confirmed that the referential coherence of discourse can be quantitatively measured with the enhanced MGCM in both Japanese and English. We will try to verify the enhanced MGCM in the other corpora of various languages or genres.

7. References

H. Alshawi. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge, England.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225, June.

K. Hasida. 1996. Issues in Communication Game. In *Proc. of COLING'96*, pages 531–536.

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

S. Lappin and H.J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

M. Poesio, R. Stevenson, B. Di Eugenio, and R. Vieira. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

S. Shiramatsu, K. Komatani, T. Miyata, K. Hasida, and H.G. Okuno. 2005. Empirical Verification of Meaning-Game-based Generalization of Centering Theory with Large Japanese Corpus. In *Proc. of PACLIC19*, pages 199–210, December.

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