# Exploring the Relationship between Semantic Spaces and Semantic Relations

**Akira Utsumi**

Department of Systems Engineering
The University of Electro-Communications
1-5-1 Chofugaoka, Chofushi, Tokyo 182-8585, Japan
utsumi@se.uec.ac.jp

**Abstract**

This study examines the relationship between two kinds of semantic spaces — i.e., spaces based on term frequency (tf) and word cooccurrence frequency (co) — and four semantic relations — i.e., synonymy, coordination, superordination, and collocation — by comparing, for each semantic relation, the performance of two semantic spaces in predicting word association. The simulation experiment demonstrates that the tf-based spaces perform better in predicting word association based on the syntagmatic relation (i.e., superordination and collocation), while the co-based semantic spaces are suited for predicting word association based on the paradigmatic relation (i.e., synonymy and coordination). In addition, the co-based space with a larger context size yields better performance for the syntagmatic relation, while the co-based space with a smaller context size tends to show better performance for the paradigmatic relation. These results indicate that different semantic spaces can be used depending on what kind of semantic relatedness should be computed.

## 1. Introduction

Recent research effort in computational lexical semantics has been directed at a semantic space model (Landauer et al., 2007; Padó and Lapata, 2007; Schütze, 1998), a corpus-based method for acquiring and representing the meaning of words. Semantic space models are computationally efficient as a way of representing meanings of words, because they take much less time and less effort to construct meaning representation and they can provide a more fine-grained similarity measure between words than other representation methods such as thesauri (e.g., WordNet). Semantic space models are also psychologically plausible; a number of studies have shown that vector-based representation achieves remarkably good performance for simulating human verbal behavior (Bullinaria and Levy, 2007; Landauer et al., 2007).

Semantic spaces (or word vectors) are constructed from large bodies of text by observing distributional statistics of word occurrence. A number of methods have been proposed for generating semantic spaces. Latent semantic analysis (LSA) is the most well-known method that uses the frequency of words (i.e., term frequency) in a fraction of documents to assess the coordinates of word vectors. Another popular method is based on the frequency of word cooccurrence within a "window" spanning some number of words (Bullinaria and Levy, 2007; Schütze, 1998).

However, despite the fact that there are different kinds of similarity between words, or different semantic relations underlying word similarity such as synonym and antonym, little has been known about the relationship between semantic spaces (or methods for constructing semantic spaces) and semantic relations. It is crucial to know what kinds of semantic relations can be represented by what kinds of semantic spaces.

One notable exception is the work of Sahlgren (2006), who relates Saussure's syntagmatic-paradigmatic distinction of semantic relation to different uses of context for computing word vectors. A semantic relation between two words is *syntagmatic* if they cooccur more often than would be expected by chance, and a semantic relation is *paradigmatic* if two words do not cooccur but they can substitute for one another. Sahlgren argues that syntagmatic relations can be represented by the use of term frequency in the same fraction of documents (which is referred to as *tf-based* in this paper), while paradigmatic relations can be represented by the use of term cooccurrence frequency within a context window of the same size (which is referred to as *co-based* in this paper). However, in his study this relationship is not justified directly or it is simply assumed to be true; he only examined which type of context use (i.e., tf-based or co-based) achieved better performance on each specific task such as word association and thesaurus comparison. For example, since word association is based on both syntagmatic and paradigmatic relations, one cannot derive any conclusions about the validity of his argument on the relationship between semantic relations and semantic spaces. Furthermore, Utsumi and Suzuki (2006) also examined the relationship between different semantic spaces and two kinds of similarity (i.e., taxonomic similarity and associative similarity), but their study suffers from the same problems that two kinds of similarity were examined without being completely separated and thus no direct evidence was obtained for the relationship between the semantic space and the semantic similarity.

This study, therefore, aims to examine the relationship between semantic relations and semantic spaces in a more systematic way. In particular, this study examines the relationship between four semantic relations (i.e., synonymy, coordination, superordination, and collocation, which are described in Section 2) and two kinds of semantic spaces (i.e., tf-based and co-based spaces, which are described in Section 3). For this purpose, in Section 4 the performance of two semantic spaces in predicting word association is compared for each semantic relation. The result of comparison is then presented and discussed in Sections 5 and 6.

Table 1: Four types of semantic relations

| Relation | Definition and description | Examples |
|---|---|---|
| Synonymy | Two words have identical or very similar meanings. | student – pupil, buy – purchase, hungry – starved |
| Coordination | Two words cluster together on the same level of detail. Antonyms come into this category. | desk – chair, red – green, black – white (antonymy) |
| Superordination | One word is a superordinate (i.e., hypernym) of another word. This category includes meronymy. | animal – dog, color – red, car – engine (meronymy) |
| Collocation | Two words are likely to cooccur in the text, because they form a predicate-argument structure. | rose – red, love – affair, baseball – play |

## 2. Semantic Relation

In this study, semantic relations are classified into four types, which are shown in Table 1. This classification is psychologically motivated; many empirical studies on word searches and speech disorders have revealed that mental lexicon is organized by these four semantic relations (Aitchison, 2003).

Among four relations, the synonymy relation is obviously paradigmatic since synonyms tend to not cooccur, while the collocation relation is thoroughly syntagmatic by definition. The coordination relation is in principle paradigmatic, but may have a syntagmatic nature. (For example, coordinate words of fruits such as apple, orange and grape may cooccur when we talk about favorite fruits.) On the other hand, the superordination relation is basically syntagmatic, but involves a paradigmatic nature. (For example, apples, oranges or grapes can be substituted for "fruits" in the sentence "This shop sells fruits.")

## 3. Semantic Space

Semantic spaces are constructed by representing all content words $t_i$ as $n$-dimensional vectors $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \cdots, w_{in})$. The degree of semantic similarity between any two words can be easily computed as, for example, the cosine of the angle formed by their vectors.

For tf-based spaces, $w_{ij}$ is calculated as term frequency i.e., the number of times the word $t_i$ occurs in a piece of text $s_j$ (e.g., paragraph or article). For co-based spaces, $w_{ij}$ is calculated as the number of times words $t_i$ and $t_j$ occur within the context window of the size $m$ (i.e., the number of times the word $t_j$ occurs within $m$ words around the word $t_i$).

## 4. Method

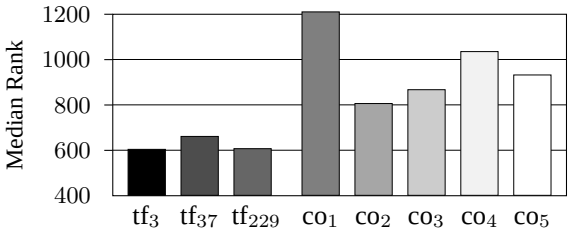
In order to examine what kind of semantic relation do semantic spaces represent, I classify word association pairs (i.e., stimulus and associate words) by their underlying semantic relations, and analysed the performance of tf-based and co-based spaces in predicting word associations for each classification. The word association pairs used in this study are top 10 associates of 201 stimulus words in a Japanese word association norm "Renso Kijunhyo" (Umemoto, 1969). These pairs are classified into four types of semantic relations listed in Table 1. Some association pairs are then removed from the analysis because their associate words (e.g., names of persons) do not or rarely occur in the corpus and thus cannot be included in the semantic spaces. As a result, four sets of word association pairs $T_{syn}$, $T_{coo}$, $T_{sup}$, and $T_{col}$ are obtained which are based on the synonymy, coordination, superordination, and collocation relations, respectively. These sets include 77, 164, 129, and 788 pairs.

Table 2: Semantic spaces used in this study

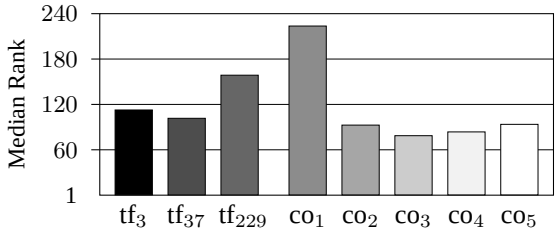
| Tf-based space | Density (%) | # of words | Dimension $n$ | Corpus size (MW[a]) |
|---|---|---|---|---|
| $\mathrm{tf}_3$ | 0.055 | 35,492 | 204,821 | 4.206 |
| $\mathrm{tf}_{37}$ | 0.136 | 34,781 | 81,548 | 4.168 |
| $\mathrm{tf}_{229}$ | 0.937 | 34,901 | 8,451 | 4.182 |
| **Co-based space** | **Density (%)** | **# of words** | **Dimension $n$** | **Window size $m$** |
| $\mathrm{co}_1$ | 0.298 | 34,781 | 34,781 | 1 |
| $\mathrm{co}_2$ | 0.589 | 34,781 | 34,781 | 2 |
| $\mathrm{co}_3$ | 0.822 | 34,781 | 34,781 | 3 |
| $\mathrm{co}_4$ | 1.012 | 34,781 | 34,781 | 4 |
| $\mathrm{co}_5$ | 1.168 | 34,781 | 34,781 | 5 |

Note. The density of semantic spaces denotes the percentage of nonzero elements of word vectors, i.e., (the number of nonzero elements) / (the number of words $\times$ $n$) $\times$ 100.
[a] MW = million words.

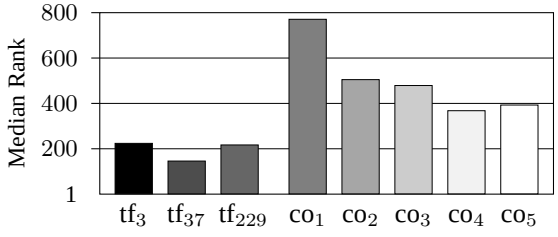
The corpus used in this study is two years' worth of Japanese Mainichi newspaper articles published in 1998 and 1999. Using this corpus, three tf-based and five co-based semantic spaces are generated, whose statistics are listed in Table 2. The tf-based spaces are generated from different sets of texts with about the same size (i.e., 4.2 million words). The set of texts for the semantic space $\mathrm{tf}_{37}$ consists of 81,548 paragraphs each of which includes 37 or more words, but the set of texts for the space $\mathrm{tf}_3$ consists of 204,821 paragraphs each of which includes 3 or more words. On the other hand, the semantic space $\mathrm{tf}_{229}$ is generated from 8,451 articles, not paragraphs, each of which includes at least 229 words, and thus it is denser than $\mathrm{tf}_3$



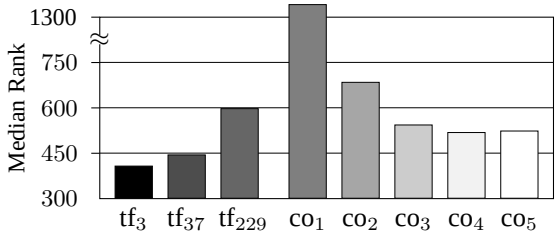

(a) Synonymy




(b) Coordination




(c) Superordination




(d) Collocation

Figure 1: Median ranks of the tf-based and co-based semantic spaces for four semantic relations.

and $\text{tf}_{37}$. All co-based spaces are generated from the set of texts for $\text{tf}_{37}$. [1]

The performance of these semantic spaces in predicting word association is measured by *the median rank* $r_{med}(T)$ of associates in a set of word association pairs $T$ (Griffiths et al., 2007).

$$r_{med}(T) = \begin{cases} r'_{\frac{|T|+1}{2}} & \text{if } |T| \text{ is odd.} \\ \frac{1}{2}(r'_{\frac{|T|}{2}} + r'_{\frac{|T|}{2}+1}) & \text{if } |T| \text{ is even.} \end{cases} \quad (1)$$

For each pair $(t_i^S, t_i^A) \in T$ of a stimulus word $t_i^S$ and an associate word $t_i^A$, the rank $r_i$ of $t_i^A$ is assessed by computing the cosine similarity between $t_i^S$ and all other words in a semantic space and sorting all words in descending order of cosine. A list of ranks $r'_1, \cdots, r'_{|T|}$ are then obtained by sorting the ranks $r_1, \cdots, r_{|T|}$ in ascending order. (Hence, for example, $r'_1$ is the minimum rank in $r_1, \cdots, r_{|T|}$.) The median rank is computed for each set of word association pairs, i.e., $T_{syn}$, $T_{coo}$, $T_{sup}$, and $T_{col}$. Smaller median ranks indicate better performance.

[1] The subscript of a tf-based space denotes the minimum number of words included in a context $s_j$ (i.e., a paragraph or an article). On the other hand, the subscript of a co-based space denotes the window size $m$.

Table 3: Comparison of my result and Sahlgren's (2006) prediction on which kind of space is more suited for representing semantic relations

| Relation | This study | Sahlgren |
|---|---|---|
| Paradigmatic relation | | |
| Synonymy | tf | co |
| Coordination | co | co |
| Syntagmatic relation | | |
| Superordination | tf | tf |
| Collocation | tf | tf |

In addition, I use a secondary measure, *recall* of associates. The recall $R_i(T)$ of a set of word association pairs $T$ is calculated as the fraction of associates $t_j^A$ which are included in the set of the top $i$ words with the highest similarity to $t_j^S$.

$$R_i(T) = \frac{|\{(t_j^S, t_j^A) \mid r_j \leq i\}|}{|T|} \quad (2)$$

Note that higher recall scores indicate better performance.

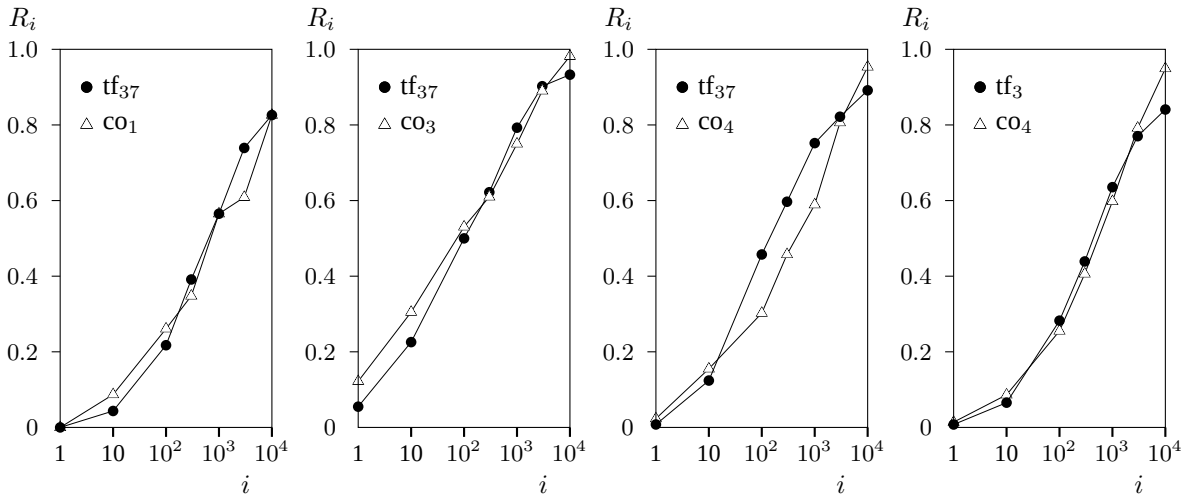
## 5. Result

Figure 1 shows the median ranks of associates computed by the eight semantic spaces for each semantic relation.

The overall result is that tf-based semantic spaces achieve better performance than co-based spaces for the synonymy, superordination, and collocation relations. Co-based semantic spaces (except $\text{co}_1$) yield better performance only for the coordination relation. The same pattern of results is obtained if the best (i.e., minimum) median rank over the eight spaces is considered; the tf-based spaces ($\text{tf}_3$ or $\text{tf}_{37}$) have the minimum median rank for the synonymy, superordination, and collocation relations, but the co-based space $\text{co}_3$ achieves the minimum median rank for the coordination relation.

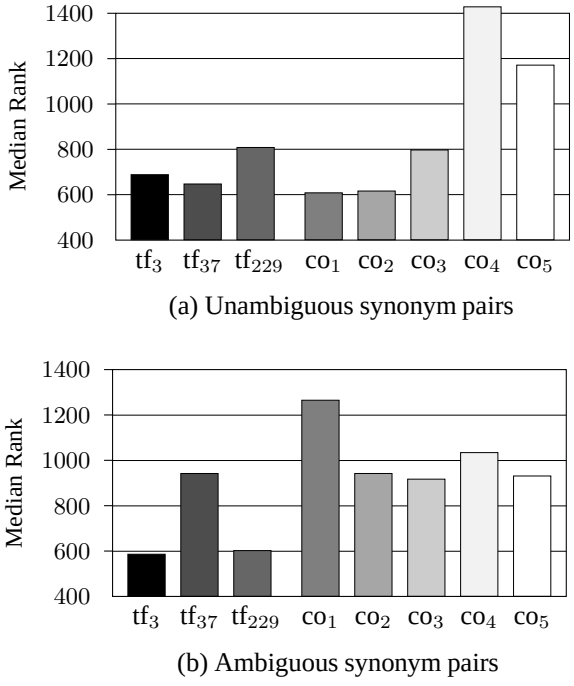
The findings on the coordination, superordination, and collocation relations are consistent with Sahlgren's (2006) argument, but the result of the synonymy relation is incompatible with his argument, as summarized in Table 3. One possible explanation of this incompatibility would be that synonyms are less likely to be substituted for each other than would be expected because all senses of the synonyms are not identical or they have different connotations of the common meaning.

To examine whether the inconsistent result of synonymy in Figure 1 is an artifact of the polysemy of synonymous words, I conducted an additional analysis in which the 77 word association pairs based on synonymy are divided into two groups — i.e., unambiguous synonym pairs and ambiguous synonym pairs — and median ranks are computed




(a) Synonym (unambiguous) (b) Coordination (c) Superordination (d) Collocation

Figure 3: Recall of four semantic relations




(a) Unambiguous synonym pairs

(b) Ambiguous synonym pairs

Figure 2: Median ranks of the tf-based and co-based semantic spaces for unambiguous and ambiguous synonyms

for each of these groups. [2] Figure 2 shows the result of this additional analysis. The result indicates that there is indeed an effect of polysemy on predicting word association based on synonymy. If the median ranks are computed for the set of unambiguous synonym pairs, the co-based spaces (in particular the spaces with a smaller window size) achieve better performance than the tf-based spaces, which is consistent with Sahlgren's (2006) prediction. On the other hand, when only association pairs of polysemous words with unshared meanigns are considered, the result of Figure 1 is replicated which contradicts Sahlgren's prediction. Hence, it is concluded that, as consistent with Sahlgren's argument, truly synonymous words, which have a considerable overlap between their multiple meanings, are better represented by co-based semantic spaces.

Concerning the relation between the co-based spaces and the paradigmatic-syntagmatic dichotomy, Peirsman et al. (2008) argued that large context windows tend to better represent syntagmatic relations, while small context windows are more appropriate for representing paradigmatic relations. Figures 1 and 2 support their argument. For the syntagmatic (i.e., superordination or collocation) relation, Figures 1(c) and (d) demonstrate that the median rank decreases (and thus the performance gets better) as the context window gets larger. On the other hand, Figure 2(a) shows that the paradigmatic (i.e., synonymy) relation is better represented by a co-based space with a smaller window size, although the performance for the coordination relation shown in Figure 1(b) does not seem to follow this pattern.

Figure 3 shows the result of an additional analysis by the measure of recall. The recall analysis gives almost the same result as the median rank analysis; the co-based space achieves higher recall for paradigmatic relations, while the tf-based space yields higher recall for syntagmatic relations, especially when $i$ is small. One interesting finding is that, regardless of semantic relations, the tf-based spaces have worse recall than the co-based spaces when $i$ is large (in particular $i = 10^4$). It is probably because the similarity between words is much more likely to be assessed as zero in the tf-based spaces due to the high data sparse-

[2] The word association pairs based on synonymy are grouped using the Japanese thesaurus "Nihongo Dai-Thesaurus" (Yamaguchi, 2006). This thesaurus consists of 1,044 basic categories, which are divided into nearly 14,000 semantic categories. Assuming that these semantic categories are distinct word senses, I count the number of senses of stimulus and associate words and assess the degree of meaning overlap of a word association pair. If more than two thirds of word senses are overlapped between stimulus and associate words, such a word association pair is classified as an unambiguous synonym pair. Note that most of the pairs classified as unambiguous have only one or two senses, which are identical between the stimulus and associate words.

Table 4: Top five associates of some stimulus words predicted by tf-based and co-based semantic spaces

| Stimulus | Associates | |
|---|---|---|
| | $tf_{37}$ | $co_3$ |
| Fish (*sakana*) | SMALL FISH (*kozakana*) | MEAT (*niku*) |
| | SWIM (*oyogu*) | VEGETABLE (*yasai*) |
| | FLOUNDER (*hirame*) | DELICIOUS (*oishii*) |
| | TROPICAL FISH (*nettaigyo*) | EAT (*taberu*) |
| | FISHERMAN (*ryoushi*) | COOKING (*ryouri*) |
| Drink (*nomu*) | TEA (*ocha*) | GET DRUNK (*you*) |
| | GLASS (*gurasu*) | EAT (*taberu*) |
| | POT (*potto*) | ALMOST DROWN (*oboreru*) |
| | PILL (*jyouzai*) | COFFEE (*kôhî*) |
| | BEER (*bîru*) | WATER (*mizu*) |

Note. Associates in ▒ and ▒ are connected to stimulus words by the superordination and coordination relations, respectively. Other associates are based on the collocation relation. Some associates are multiwords, but their original Japanese expressions are single words.

ness. However, the result of Figure 1 that for none of the semantic relations does the least sparse tf-based space $tf_{229}$ achieve the best performance indicates that reduction of sparseness by using a larger context is not necessarily an efficient way of overcoming this difficulty. Dimensionality reduction may provide a solution to this problem; LSA takes this approach by reducing the original semantic space into a much smaller space by using a technique of singular value decomposition. It would be interesting to examine the effects of dimensionality reduction on the representational power of semantic spaces, although I do not discuss this issue here.

Table 4 shows some examples of the top five associates produced by two semantic spaces $tf_{37}$ and $co_3$. The result is consistent with the finding obtained by median rank; the tf-based space $tf_{37}$ is likely to list as top associates subordinate or collocated words of the stimulus, but the co-based space tends to pick out coordinate words.

## 6. Discussion: Do Subcategories of Semantic Relations Yield Consistent Results?

Some of the four relations used in this study can be further divided into subcategories, which have been discussed in lexical semantics (Cruse, 1986). Hence, the following

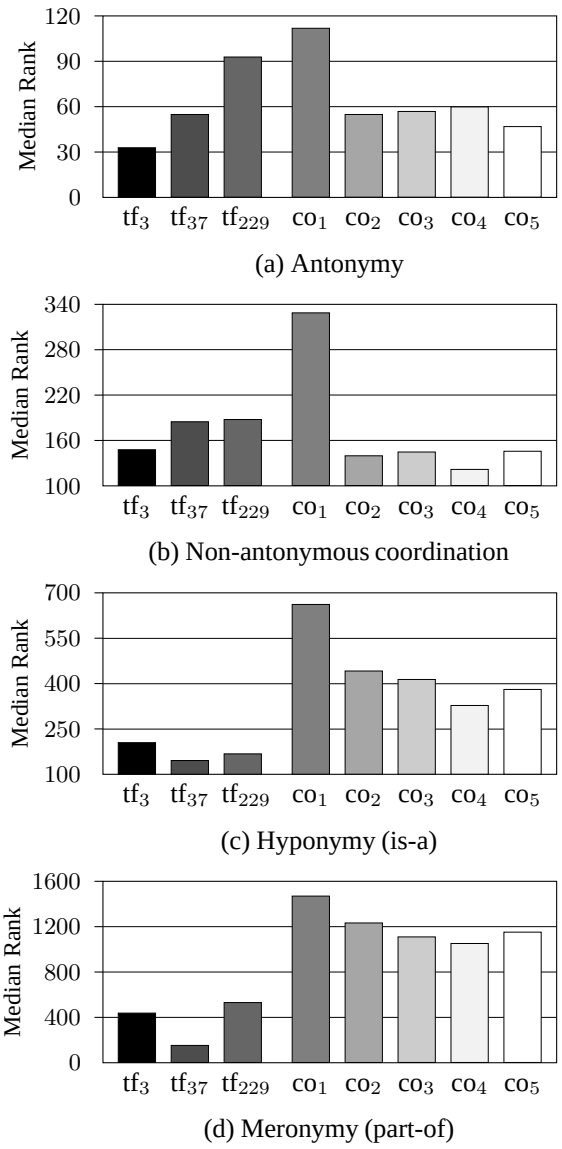



Figure 4: Median ranks of the semantic spaces for subcategories of the semantic relations

subrelations are considered and the median ranks for these subrelations are computed. [3]

- Coordination
  - Antonymy (or opposition): Two words have the opposite meaning. (e.g., black – white).
  - Non-antonymous coordination: Two words do not have the opposite meaning but cluster together on the same level. (e.g., desk – chair).
- Superordination
  - Hyponymy: One word is a kind of another word. Hyponymy is also referred to as a "is-a" relation. (e.g., animal – dog).

[3] The 164 word association pairs based on the coordination relation are divided into 66 antonymy pairs and 98 non-antonymous pairs. Similarly, the 129 word association pairs for the superordination relation contain 96 hyponymy pairs and 33 meronymy pairs.

– Meronymy: One word is a part of another word. Meronymy is also referred to as a "part-of" relation. (e.g., engine – car).

Figure 4 shows the result of the tf-based and co-based spaces for each of these subrelations. Concerning the subcategories of superordination, they show the same pattern of results that the tf-based spaces yield better performance, although the performance of the meronymy (part-of) relation is totally worse than that of the hyponymy (is-a) relations. For the subcategories of coordination, however, one noteworthy finding is obtained which is incompatible to the finding of Figure 1; the antonymy relation is better represented by the tf-based space $tf_3$ than by the co-based spaces, although the co-based performance is also very high. This finding may be due to the distinctive nature of antonymy that antonymous words are more likely to cooccur than other coordinate words; we often say "good news and bad news," "boy meets girl," "win or lose," and so on. Hence, antonymy has a syntagmatic nature and thus can be better represented by the tf-based spaces.

## 7. Concluding Remarks

In this paper, I have examined the relationship between two kinds of semantic spaces (i.e., tf-based and co-based) and four semantic relations (i.e., synonymy, coordination, superordination, and collocation) in a systematic way, and obtained the following findings.

- The tf-based and co-based semantic spaces better represent different kind of semantic relations; A co-based space is suited for representing the paradigmatic relation, while a tf-based space is appropriate for the syntagmatic relation. This finding is consistent with Sahlgren's (2006) claim.
- Computing the semantic relatedness between synonymous words is affected by their polysemy. When synonymous words are highly polysemous and thus they are not truly synonyms, co-based spaces are less appropriate for computing the semantic similarity.
- Antonymy is better judged by the tf-based spaces, in contrast to non-antonymous coordination, which is better represented by the co-based spaces. This may be because antonymy is more syntagmatic in nature.
- The co-based space with a larger context size yields better performance in computing the semantic relatedness based on the syntagmatic relation, while the co-based space with a smaller context size tends to show better performance for the paradigmatic relation. This finding is consistent with Peirsman et al.'s (2008) claim.

These findings indicate that different semantic spaces can be used depending on what kind of semantic similarity should be computed.

It would be vital for further research to examine whether these findings are generalised to other space generation methods including a topic model (Griffiths et al., 2007), other languages, other similarity measure, and other kinds of corpus, as well as to develop a versatile method for representing all kinds of semantic relations. Another important topic that should be addressed is to examine the effectiveness of dimensionality reduction, and the relationship between dimensionality reduction techniques and semantic relations.

## 8. Acknowledgment

This study was supported by a Grant-in-Aid for Scientific Research(C) (No.20500234) from Japan Society for the Promotion of Science.

## 9. References

Aitchison, J. (2003). *Words in the Mind: An Introduction to the Mental Lexicon, 3rd Edition*. Oxford, Basil Blackwell.

Bullinaria, J.A. and Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Cruse, D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Griffiths, T.L., Steyvers, M., and Tenenbaum, J.B. (2007). Topics in semantic representation. *Psychological Review*, 114:211–244.

Landauer, T.K., McNamara, D.S., Dennis, S., and Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Peirsman, Y., Heylen, K., and Geeraerts, D. (2008). Size matters: Tight and loose context definitions in English word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41.

Sahlgren, M. (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces*. Ph.D. thesis, Stockholm University.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Umemoto, T. (1969). *Renso Kijunhyo (Free Association Norm)*. Tokyo Daigaku Shuppankai, Tokyo.

Utsumi, A. and Suzuki, D. (2006). Word vectors and two kinds of similarity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006) Main Conference Poster Sessions*, pages 858–865.

Yamaguchi, T. (2006). *Nihongo Dai-Thesaurus CD-ROM (Japanese Thesaurus)*. Taishukan Shoten, Tokyo.