Semantic Analysis of Abstract Nouns to Compile a Thesaurus of Adjectives

Kyoko Kanzaki[1],   Qing Ma[2],   Eiko Yamamoto[1] and Hitoshi Isahara[1]

[1]Computational Linguistics Group,

National Institute of Information and Communications Technology

3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto, Japan,

{kanzaki, eiko, isahara} @nict.go.jp

[2] Faculty of Science and Technology

Ryukoku University

Seta, Otsu,520-2194, Japan

qma@math.ryukoku.ac.jp

Abstract

Aiming to compile a thesaurus of adjectives, we discuss how to extract abstract nouns categorizing adjectives, clarify the semantic and syntactic functions of these abstract nouns, and manually evaluate the capability to extract the "instance-category" relations. We focused on some Japanese syntactic structures and utilized possibility of omission of abstract noun to decide whether or not a semantic relation between an adjective and an abstract noun is an "instance-category" relation. For 63% of the adjectives (57 groups/90 groups) in our experiments, our extracted categories were found to be most suitable. For 22 % of the adjectives (20/90), the categories in the EDR lexicon were found to be most suitable. For 14% of the adjectives (13/90), neither our extracted categories nor those in EDR were found to be suitable, or examinees' own categories were considered to be more suitable. From our experimental results, we found that the correspondence between a group of adjectives and their category name was more suitable in our method than in the EDR lexicon.

1. Introduction

A thesaurus is a classification of words according to an organization of concepts categorizing words semantically. There already exist several kinds of thesauri, such as WordNet in English, the EDR concept dictionary in Japanese and English, and so on. Since these thesauri have been primarily constructed manually, there may be some defects in their organizational structure, making it crucial to verify their accuracy.

There are at least two ways to evaluate the usefulness of a thesaurus. One is to apply it to practical tasks, such as information retrieval and sentence understanding. Another is to construct hierarchies from real data, e.g., a huge corpus, and compare them to the hierarchies in a hand-coded thesaurus.

So far, several approaches to the automatic extraction of thesauri or ontology from huge corpora have been undertaken. One is the categorization of words into semantic classes by calculating the distributions of words in the corpus using syntactic and/or surface patterns (Hindle 1990, Hatzivassiloglou and MacKeown 1993, Lin and Pantel 2001, Walde and Brew 2002, Boleda et al. 2004, etc.), and another is the statistical extraction of semantic relations of words, such as hypernym-hyponym relations or part-whole relations, from corpora using syntactic patterns (Hearst 1992, Caraballo 1999, Berland and Charniak 2000, Rydin 2002, and Pantel and Ravichandran 2004, etc.). Most of these research projects use linguistic patterns to extract hypernyms of nouns.

However, though these linguistic patterns are useful to extract hypernym-hyponym relations between nouns or relations between a noun and a verb, it is difficult to extract from data semantic categories of other parts of speech, such as adjectives, using the same kind of linguistic clues. We need to determine linguistic clues which are suited to the extraction of semantic categories of other parts of speech.

Aiming to compile a thesaurus of adjectives, we discuss how to extract abstract nouns categorizing adjectives, clarify the semantic and syntactic functions of these abstract nouns, and manually evaluate the capability to extract the "instance-category" relations. Note that we treat each abstract noun not as a kind of noun but as a category of adjectives.

2. How to Extract Abstract Nouns Categorizing Adjectives from Corpora

Let's discuss the Japanese syntactic structure "Noun1 *wa* Noun2 *ga* Adj.," in which "Noun1 *wa*" refers to a topic and "Noun2 *ga*" refers to a subject. According to the Japanese linguist Takahashi (1975), in the case of "*Yagi wa seishitsu ga otonashii* (The nature of a goat is gentle)," "*seishitsu* (nature)" (N2) is a super-ordinate concept of "*otonashii* (gentle)" (Adj), and conversely, the adjective "*otonashii* (gentle)" includes the meaning of the abstract noun "*seishitsu* (nature)." In this sentence, the abstract noun "*seishitsu* (nature)" can be omitted without changing the meaning of this sentence, i.e., the meanings of "*Yagi wa otonashii* (A goat is gentle)," and "*Yagi wa*

seishitsu ga otonashii (The nature of a goat is gentle)" are the same. Takahashi concluded that in this type of semantic relation between an adjective and a noun, the noun "*seishitsu* (nature)" was considered to be a kind of super-ordinate noun of the adjective "*otonashii* (gentle)."

On the other hand, "*Zou wa hana ga nagai* (The trunk of an elephant is long)" has the same syntactic structure as the previous example, i.e., "Noun1 *wa* Noun2 *ga* Adj." However, the noun "*hana* (trunk)" cannot be omitted without changing a meaning of the sentence. "*Zou wa nagai* (An elephant is long)" is incorrect or ambiguous. The meaning of the word "*nagai* (long)" is not included in a meaning of "*hana* (trunk)." In the semantic relation between an adjective and the noun "*hana* (trunk)," "*hana* (trunk)" is not considered as a kind of super-ordinate noun of the adjective "*nagai* (long)."

We basically utilize this feature, i.e., the possibility of the omission of abstract nouns, to decide whether or not a semantic relation between an adjective and an abstract noun is an "instance-category" relation.

According to Isahara and Kanzaki (1999), there are two patterns in which the relation between an adjective and its head noun is an "instance-category" relation in certain situations. One is the "Noun1 *ga* Adj (Noun1 is Adj)" pattern, for example "*seishitsu ga otonashii* (the nature is gentle)" and the other is the "Adj + Noun" pattern which cannot be paraphrased into a predicative relation, for example, "*kanashii kimochi* (sad feeling)."

However, all examples of such "instance-category" relations are not always extracted by gathering only the syntactic patterns such as "Noun1 *ga* Adj (Noun1 is Adj)" and "Adj + Noun."

It is necessary, therefore, to narrow the patterns. The procedure we used to do so is as follows:

Step1) Extract from the corpora all nouns which are preceded by the Japanese expression "*to iu*" which is similar to "that" or "of" in English. "*To iu* + noun (noun that/of ...)" is a typical Japanese expression which introduces some information about the referent of the noun, such as apposition. Therefore, nouns found in this pattern may have their content elucidated by means of their modifiers.

Step2) Extract all adjectives modifying the nouns extracted in step 1 from the corpora.

NB: the relationships between adjectives and their modifying nouns extracted here include not only "instance-category" relations, but also other various relations.

Step3) Extract "instance-category" relations between adjectives and nouns from data gathered in step2.

In Step3, we used syntactic patterns such as "Noun1 *wa* Noun2 *ga* Adj" and "Adj + Noun2_*no*+ Noun1" in order to determine an "instance-category" relation. "Noun1" is a concrete noun representing a topic or a subject, and "Noun2" is elucidated by its modifier, an adjective. "*No*" is a marker of adnominal usage of a noun. Among the data gathered in Step2, we choose examples in which Noun2 can be omitted without changing the meaning of the original sentence or phrase. If Noun2 can be omitted, Noun2 may be an abstract concept of the modifying

adjective. For example,

Yasashii seishitsu no yagi
 Adj.   Noun2 "no"-marker Noun1
 (a goat having a gentle nature)

In this case, since we can say "*yasashii yagi* (a gentle goat)," "*seishitsu* (nature)" is identified as a kind of a category of "*yasashii* (gentle)."

The data that we extracted is as follows,

<u>*KIMOCHI*</u> (feeling): *ureshii* (glad), *kanashii* (sad), *shiawasena* (happy), …

<u>*OMOI*</u> (thought): *ureshii* (glad), *tanoshii* (pleased), *hokorashii* (proud), …

<u>*KANTEN*</u> (viewpoint): *igakutekina* (medical), *rekishitekina* (historical), ...

We extracted abstract nouns from the Mainichi Shinbun newspaper, and adjectives co-occurring with the abstract nouns in the above mentioned manner from 100 novels, 100 essays, and 42 years' worth of newspaper articles, including 11 years' worth of Mainichi Shinbun articles, 10 years' worth of Nihon Keizai Shinbun (Japanese economic newspaper) articles, 7 years' worth of Sangyoukinyuryuutusu shinbun (Japanese economic newspaper) articles, and 14 years worth of Yomiuri Shinbun articles. We have now gathered about 365 of such abstract nouns.

3. Semantic Function of Abstract Nouns

In step3 we chose examples depending on the possibility of omitting abstract nouns in order to determine whether or not a semantic relation between an adjective and an abstract noun is an "instance-category" relation. In this section we analyze examples in which abstract nouns can and cannot be omitted.

In the case that an abstract noun can be omitted, it categorizes an adjectival meaning, while in the case that an abstract noun cannot be omitted, it adds a grammatical function and a shade of meaning to an adjectival meaning. Following, we describe such cases in detail.

A) In the case that abstract nouns can be omitted: categorization of adjectival meaning

For example, "color" is a category of "red, blue, white," and so on. Some examples are given below.

concept names categorizing adjectives:
                              instances of adjectives

Kanshoku (feel):
    yawarakai (soft), *katai* (solid), *shimeppoi* (damp),
    namerakana (smooth), …
Kansei (sensibility):
    naibuna (naive), *sensaina* (delicate),
    shinayakana (exquisite),
Katachi (shape):
    shikakui (square), *marui* (round), *hiratai* (flat),
    kyokusentekina (sweeping), …

Some abstract nouns can represent a small range of connotation and a wider range of denotation. Since this type of abstract noun is an expression of formalization, the meaning of an abstract noun is more transparent. In this case, an abstract noun is an upper-level concept. Some examples are given below.

Jotai (condition):
    fuanteina jotai (unstable condition)
Katachi (the form): *shitashimiyasui katachi de …*
                    (in a form which is easily familiar …)

In the following example "*jotai* (condition")" can be ommited.

kokkai   wa fuanteina   joutai   da.
(national assembly)   (unstable) (condition) (auxiliary
                                     indicating predication)
(The national assembly is now in unstable condition.)

In this case, "*jotai* (condition)" can be omitted if the usage of the adjective "*fuanteina*" is changed from adnominal usage to predicative usage, that is, "*fuanteina*"(in adnominal usage) → "*fuanteida*" (in predicative usage).

(*kokkai wa*)
    fuanteina   joutai   da
    (unstable) (condition)   (auxiliary indicating
                             predication)

→ (kokkai wa) *fuanteida*.
                (be unstable)
    (The national assembly is unstable.)

In this example, "*jotai* (condition)" is transparent.

In other examples, "an adjective in adnominal usage + an abstract noun (head noun) + *de* (auxiliary indicating adverbial usage)" can be changed into the adjective in adverbial usage without changing the phrasal or sentential meaning. In the following example, we find that the meaning of "*katachi* (form)" is transparent.

Kono   mondai   ga   namanamashii   katachi de
(This) (problem) s-marker   (vivid)   (in the form)
                    arawaretekuru.
                    (appear)
    (The problem is appearing in a vivid form.)

"De" following "katachi (form)" is an auxiliary for a marker of adverbial usage. In this example, "*namanamashii katachi de*" can be changed into "*namanamashiku*," which is the adverbial usage of "*namanamashii*," without changing its meaning.

Kono mondai ga namanamashiku arawaretekuru.
                (vividly)
    (The problem is appearing vividly.)

B) In the case that abstract nouns cannot be omitted: an abstract noun adds a grammatical function and a shade of meaning to an adjectival meaning.

B-1) Abstract nouns have the grammatical function of a "particle" and add subtle meanings to adjectives.

In the following examples, the abstract nouns represent a time period of a given situation or the state of being in a certain situation: *uchi* (while), *naka* (in):

    atsui *uchi* (while it is hot),
    kensou no naka (in the noise).

B-2) Abstract nouns have the grammatical function of a "connection particle" or "particle representing a mood."

1)   Some abstract nouns with the grammatical function of a connection particle are representing the degree of something:

hodo (The ADJ_er .., the ADJ_er…), …
"*Denryu ga ookii hodo, jikai ga kyoryoku ni naru*
(The higher the current, the stronger the magnetic field)"

2)   An abstract noun is representing a euphemism, so it adds a mood to the phrasal or sentential meaning that the adjective is expressing:

Atari (around, "…, say, …"), …

Hagire no ii atari   ga   ninki no himitsu
(liveliness) (, say,) subject marker (popularity) (secret)
(The secret of his popularity may be that he is, say, lively.)

4.   Verification

Among the "instance-category" relations that we extracted according to the procedures outlined in sections 2 and 3, we manually evaluated abstract nouns that we regarded as categories of adjectives by comparing them with a Japanese thesaurus, the EDR lexicon.

4.1. Experimental procedure

We prepared sets of groups of adjectives extracted in sections 2 and 3 and abstract nouns categorizing adjectives. We recruited 5 examinees for these experiments.

1) We showed examinees a group of adjectives and had them determine its category by themselves.

Example)
Question:
    Please define a category of adjectives shown in this list.

Adjectives: ***hagireno_warui***   (**inarticulate**), ***shitatarazuna*** (**babble**), ***togetogeshii*** (**acrid**), ***netsuppoi*** (**vehement**), ***kuchihabattai*** (**shoot off one's mouth**), ***iyamina*** (**bitteness**)
Your category: ***

2) For each group of adjectives, we had examinees judge the suitability of categories corresponding to the group of adjectives by comparing three categories, i.e., their own category created in step1, our extracted category, and the category in the EDR lexicon. Also, we had them consider whether or not the group of adjectives needs a category.

Question:

Please choose a suitable category among three. Note that the third category is your defining category.

Adjectives: *hagireno_warui* (Inarticulate), *shitatarazuna* (babble), *togetogeshii* (acrid), *netsuppoi* (vehement), *kuchihabattai* (shoot off one's mouth), *iyamina* (bitteness)
(1) *Iikata* (way of saying)
(2) *Taido ya seikaku no atai* (a value of one's attitude or character)
(3) *Iikata* (way of saying)

4.2. Experimental data

For 18 groups of adjectives that we randomly chose in our data, 5 persons judged the suitability of the three categories. Therefore, we investigated 90 (18 x 5) questions in total.The data used in our experiment is shown below. The numbers, (1) and (2), in this list show our extracted category and EDR category respectively.

ID 1	*hagireno_warui* (inarticulate), *shitatarazuna* (babble), *togetogeshii* (acrid), *netsuppoi* (vehement), *kuchihabattai* (shoot off one's mouth), *iyamina* (bitteness), …		
(1)	*Iikata* (way of saying)	(2)	*Taido ya seikaku no atai* (a value of one's attitude or character)
ID 2	*sugenai* (inhospitable), *nanigenai* (casual), *sokkenai* (brief), *araarashii* (rude), *ayashii* (questionable), *yasashii* (kind), …		
(1)	*Soburi* (behavior)	(2)	*Taido ya seikaku no atai* (a value of one's attitude or character)
ID 3	*kisokutadashii* (orderly), *sugasugashii* (breezy), *furumekashii* (ancient-looking), *sappuukeina* (bleak), *kansona* (simple), *shissona* (simple), …		
(1)	*Tatazumai* (appearance)	(2)	NONE
ID 4	*atatakai* (warm), *uttoushii* (gloomy), *tsuyunoyouna* (rainy season-like), *natsurashii* (summerlike), *odayakana* (clement), …		
(1)	*Tenki* (climate)	(2)	*Aruchiikino cyoukikan ni wataru kishou no joutai* (The condition of the weather which lasts for a long time in a certain area)
ID 5	*kokochiyoi* (comfortable)		
(1)	*Yasuragi* (comfort)	(2)	*Kaiteki de kibun ga yoi sama* (The comfortable state which feels good)
ID 6	*hanabanashii* (steller), *ikinonagai* (prolonged), *kattatsuna* (lively), *enerugisshuna* (energetic), *kakan'na* (valiant), …		
(1)	*Katsuyaku* (fine showings)	(2)	*Hitoya monogoto ga yoi inshou wo ataeteiru* (a person and things give a good impression to someone)

ID 7	*pawafuruna* (powerful), *wakawakashii* (youthful), *takumashii* (strong spirit), *seishin'na* (fresh), …		
(1)	*Katsuryoku* (energy)	(2)	NONE
ID 8	*surudoi* (keen)		
(1)	*Kan* (flair)	(2)	*Kankaku ga kiwamete eibin'na sama* (the state that a sense is very keen)
ID 9	*atatakai* (warm), *yawarakai* (soft), *katai* (solid), *shimeppoi* (dank), …		
(1)	*Kanshoku* (feeling)	(2)	NONE
ID 10	*ryuchouna* (fluent), *tadotadoshii* (bumbleheaded), *gyougyoushii* (orotund), …		
(1)	*Hanashikata* (way of speech)	(2)	*Hyoujou dousa no youtai* (the manner of an expression and a movement)
ID 11	*aimaina* (ambiguous), *koikina* (smartish), *in'utsuna* (sullen), *jiyuukattatsuna* (fluid), *nodokana* (idyllic)		
(1)	*Fun'iki* (atmosphere)	(2)	NONE
ID 12	*sugoi* (like crazy), *susamajii* (furious), *pawafuruna* (powerful), *attoutekina* (overwhelming), ..		
(1)	*Hakuryoku* (power)	(2)	*Jishou ni kakawaru hindo, jikan, sokudo, teido no youtai* (The aspect of the frequency concerned with the matter, the time, the speed and the degree)
ID 13	*komayakana* (paying close attention to), *saishin'na* (solicitous), *shinsetsuna* (hospitality), *yasashii* (tender)		
(1)	*Hairyo* (care)	(2)	NONE
ID 14	*ken'akuna* (stormy), *shinmitsuna* (cordial), *shitashii* (close), *kiyasui* (friendly), …		
(1)	*Naka* (relationship)	(2)	*Hito doushi no sinriteki na chikasa* (nearness of the mental distance between persons)
ID 15	*ketatamashii* (shrill), *urusai* (noisy)		
(1)	*Souon* (noise)	(2)	*Oto ya koe nado ga ookii sama* (A state that the sound and voice is loud.)
ID 16	*asekurai* (sweaty), *namagusai* (smelly), *kogekusai* (scorched flavor), ….		
(1)	*Shuuki* (off-flavor)	(2)	*Nioino atai* (a vlue of odor)
ID 17	*toritomenonai* (disjointed), *sokkyoutekin* (ad lib), *toppina* (outrageous), …		
(1)	*Omoitsuki* (idea)	(2)	NONE
ID 18	*kagayakuyouna* (sparkling), *migaitayouna* (polished - like), *tsuyayakana* (shiny), …		
(1)	*Koutaku* (glaze)	(2)	*Kin'kira kagayaiteiru sama* (A state that something is glitting)

Table 1: Our data used for our experiment

4.3. Examinees

Five examinees were chosen from a group of linguists, persons engaged in Japanese education, and persons in NLP.

4.4. Experimental Results

The suitability of categories among the three methods was calculated as follows:

For 63% of the adjectives (57 groups/90 groups), our extracted categories were found to be most suitable.

For 22 % of the adjectives (20/90), the categories in the EDR lexicon were found to be most suitable.

For 14% of the adjectives (13/90), neither our extracted categories nor those in EDR were found to be suitable, or examinees' own categories were considered to be more suitable.
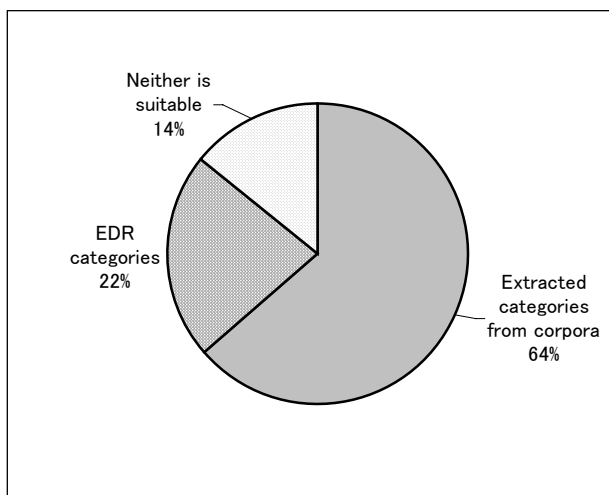

Fig1. Suitability of categories among the three methods.

As Figure 2 shows, we itemized the suitability of categories among the three methods.
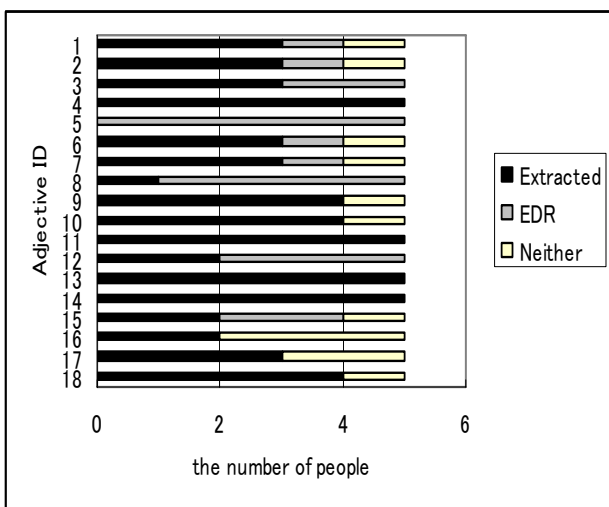

Fig.2 Breakdown of the suitability of categories among the three methods.

Categories that all examinees (5 persons) agreed were:

In our extracted categories

ID 4, "*tenki* (climate)" ,

ID 13, "*hairyo* (care)",

ID 14, "*naka* (relationship)"

In EDR categories

"*Kaiteki de kibun ga yoi sama*
(The comfortable state which feels good, ID 5)".

Categories that 4 persons agreed were:

In our extracted categories

ID 9, "*kanshoku* (feeling)"

ID 10, "*Hanashikata* (way of speech)"

ID 11, "*Fun'iki* (atmosphere)"

ID 18, "*Koutaku* (glaze)"

In EDR categories

ID 8, "*Kankaku ga kiwamete eibin'na sama*
(the state that a sense is very keen)"

According to examinees' comments, in cases in which a category was not found in EDR, most examinees considered that a category was needed for the group of adjectives.

Other comments noted that it was sometimes difficult for examinees to determine the most suitable category name because the category names were too abstract (both in our method and EDR).

From our experimental results, we found that the correspondence between a group of adjectives and their category name was more suitable in our method than in the EDR lexicon.

5. Conclusion

We determined some linguistic clues for the extraction of semantic categories of adjectives.

Aiming to compile a thesaurus of adjectives, we discussed how to extract abstract nouns categorizing adjectives, clarified the semantic and syntactic functions of these abstract nouns, and evaluated human capability to extract "instance-category" relations.

In our experiment, we found that our extracted categories were more suitable than those in the EDR lexicon. As our research continues, we will try to further extract suitable categories of adjectives.

References

Matthew Berland, and Eugene Charniak. 2000. "Finding Parts in Very Large Corpora" In *Proceedings of 38th Annual Meeting of the Association for Computatoinal Linguistices (ACL)*. pp.57-64.

Gemma Boleda, Toni Badia, and Eloi Batlle. 2004. "Acquisition of Semantic Classes for Adjectives from Distributional Evidence." In *Proceedings of 20th International Conference on Computational Linguistics (COLING), Volume 2*. pp.1119-1125.

Sharon A. Caraballo. 1999. "Automatic construction of a hypernym-labeled noun hierarchy from text." In *Proceedings of 37th Annual Meeting of the Association for Computatoinal Linguistices (ACL).* pp.120-126.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning." In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics (ACL).* pp.172-182.

Marti A. Hearst. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Proceedings of 14th International Conference on Computational Linguistics (COLING).* pp.539-545.

Donald Hindle. 1990. "Noun Classification From Predicate-Argument Structures." In *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics (ACL).* pp.268-275.

Hitoshi Isahara and Kyoko Kanzaki. 1999. "Lexical semantics to disambiguate polysemous phenomena of Japanese adnominal constituents." In *Proceedings of 37st Annual Meeting of the Association for Computational Linguistics (ACL).* pp. 489-496.

Dekang Lin and Patrik Pantel. 2002. "Concept Discovery from Text." In *Proceedings of 19th International Conference on Computational Linguistics(COLING).* pp.768-774.

Patrick Pantel and Deepak Ravichandran. 2004. "Automatically Labeling Semantic Classes." In *Proceedings of Human Language Technology / North American Chapter of the Association for Computational Linguistics (HLT/NAACL-04).* pp. 321-328.

Sara Rydin. 2002. "Building a hyponymy lexicon with hierarchcal structure." 2002. In *Proceedings of the workshop on Unsupervised Lexical Acquisition (SIGLEX).* pp.26-33.

Sabine Schulte im Walde and Chris Brew. 2002. "Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information." In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL).* pp.223-230.

Taro Takahashi. 1975. "A various phase related to the part-whole relation investigated in the sentence." *Studies in the Japanese language* 103: The Society of Japanese Linguistics. pp. 1-16.