Qualia structures and their impact on the concrete noun categorization task

Katrenko, S.; Adriaans, P.

Published in:
Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: Bridging the gap between semantic theory and computational simulations

Qualia Structures and their Impact on the Concrete Noun Categorization Task

Sophia Katrenko
Informatics Institute
University of Amsterdam
the Netherlands
katrenko@science.uva.nl

Pieter Adriaans
Informatics Institute
University of Amsterdam
the Netherlands
pietera@science.uva.nl

Abstract

Automatic acquisition of qualia structures is one of the directions in information extraction that has received a great attention lately. We consider such information as a possible input for the word-space models and investigate its impact on the categorization task. We show that the results of the categorization are mostly influenced by the formal role while the other roles have not contributed discriminative features for this task. The best results on 3-way clustering are achieved by using the formal role alone (entropy 0.00, purity 1.00), the best performance on 6-way clustering is yielded by a combination of the formal and the agentive roles (entropy 0.09, purity 0.91).

1 Introduction

Computational models of semantic similarity have been used for some decades already with various modifications (Sahlgren, 2006). In this paper, we investigate qualia structures and their impact on the quality of the word-space models. Automatic acquisition of qualia structures has received a great attention lately resulting in several methods which use either existing corpora or the Web (Cimiano and Wenderoth, 2007; Yamada et al., 2007). We build on the work reported in the literature and aim to test how suitable the results of automatic qualia extraction are for the word-space models. We approach a seemingly simple task of the concrete noun categorization. Previous research has shown that when humans are asked to provide qualia elements per role for a list of nouns, concrete nouns lead to the highest agreement. The words with the lowest agreement are abstract notions (Cimiano and Wenderoth, 2007). Naturally, a question arises of what information would be captured by the word-space models if qualia elements are used.

This paper is organized as follows. Section II presents some relevant information on word-space models and their modifications. Section III gives a brief overview of the Generative Lexicon Theory. Then, we describe a method used for an automatic qualia structure acquisition. We proceed with an experimental part by discussing results and analyzing errors.

2 Word-Space Models

Underlying idea behind the word-space models lies in the semantic similarity of words. In particular, if two words are similar, they have to be close in the word space which led to so called geometric metaphor. In his dissertation, Sahlgren (2006) discusses different ways of constructing such word spaces. One possible solution is to take into account word co-occurrences, the other would be using a limited number of semantic features. While the former method may result in a high-dimensional space containing redundant information, the latter may be too restrictive. The main concern about a list of features is how they can be defined and what kind of features are sufficient for a given task. Sahlgren (2006) argues that the word-space models have to be considered together with a task they are used for. He highlights differences between the word-space models based on paradigmatic and syntagmatic notions and shows that both models can be effectively used. On the task of human association norm, word spaces produced by using syntagmatic information seem to have a higher degree of corelation with a norm, while paradigmatic word-spaces yield better results on the synonymy test.

3 Generative Lexicon Theory

In the semantic theory of Generative Lexicon, Pustejovsky (2001) proposes to describe lexical expressions by using four representation levels, argument structure, event structure, qualia structure, and lexical inheritance structure. For the work presented here, qualia structure is of the most interest. Qualia structure use defined by the following roles:

- *formal* - information that allows to distinguish a given objects from others, such as superclass
- *constitutive* - an object's parts
- *telic* - a purpose of an object; what it is used for
- *agentive* - origin of an object, "how it came into being"

While discussing natural kinds and artifacts, Pustejovsky (2001) argues that a distinction between these two categories can be drawn by employing a notion of intentionality. In other words, it should be reflected in the telic and agentive roles. If no intentionality is involved, such words are natural types. On the contrary, artifacts are identified by the telic and agentive roles.

4 Automatic Acquisition of Qualia Structures

After the theory of Generative Lexicon has been proposed, various researchers put it in practice. For instance, Lenci (2000) considered it for designing ontologies on example of SIMPLE. Qualia structure is used here to formally represent a core of the lexicons. In a nutshell, the SIMPLE model is more complex and besides qualia structure includes such information as argument structure for semantic units, selectional restrictions of the arguments, collocations and other. The Generative Lexicon theory was also used for different languages. For instance, Zavaglia and Greghi (2003) employ it to analyze homonyms in Portuguese.

Another interesting and useful aspect of qualia structure acquisition is automatic qualia extraction. Recently, Yamada et al. (2007) presented a method on the telic role acquisition from corpus data. A motivation behind the telic role was that there are already approaches to capture formal or constitutive information, while there is less attention to the function extraction. A method of Yamada et al. (2007) is fully supervised and requires a human effort to annotate the data.

Contrary to the work reported in (Yamada et al., 2007), Cimiano and Wenderoth (2007) proposed several hand-written patterns to extract qualia information. Such patterns were constructed in the iterative process and only the best were retained. Further, the authors used various ranking measures to filter out the extracted terms.

We start with the qualia information acquisition by adopting Cimiano and Wenderoth's (2007) approach. For each role, there is a number of patterns which might be used to obtain qualia information. Table 1 contains a list of the patterns per role which have been proposed by Cimiano and Wenderoth (2007). All patterns are accompanied by the parts of speech tags.[1] The patterns for the *formal* role are well-known Hearst patterns (Hearst, 1992) and patterns for the other roles were acquired manually.

In Table 1 x stands for a seed in singular (e.g., *lion*, *hammer*) and p for a noun in plural (e.g., *lions*, *hammers*). For pluralia tantum nouns only a corresponding subset of patterns is used. In addition, we employ a wildcard which stands for one word (a verb, as it can be seen in agentive patterns).

5 Experiments

The data set used in our experiments consists of 44 words which fall in several categories, depending on the granularity. On the most general level, they can be divided in two groups, *natural kind* and *artifact*. Further, *natural* group includes such categories as *vegetable* and *animal*. On the most specific level, the data set represents the following 6 categories: *green*, *fruitTree*, *bird*, *groundAnimal*, *tool*, and *ve-*

[1] the following categories are used : nouns in singular (NN), nouns in plural NNP, conjunctions (CC), determiners (DET), adjectives (JJ), prepositions (IN)

Role	Pattern
formal	x_NN is_VBZ (a_DT|the_DT) kind_NN of_IN x_NN is_VBZ x_NN and_CC other_JJ x_NN or_CC other_JJ such_JJ as_IN p_NNP *,*(* especially_RB p_NNP *,*(* including_VVG p_NNP
telic	purpose_NN of_IN (a_DT)* x_NN is_VBZ purpose_NN of_IN p_NNP is_VBZ (a_DT|the_DT)* x_NN is_VBZ used_VVN to_TO p_NNP are_VBP used_VVN to_TO
constitutive	(a_DT|the_DT)* x_NN is_VBZ made_VVN (up_RP)*of_IN (a_DT|the_DT)* x_NN comprises_VVZ (a_DT|the_DT)* x_NN consists_VVZ of_IN p_NNP are_VBP made_VVN (up_RP)*of_IN p_NNP comprise_VVP p_NNP consist_VVP of_IN
agentive	to_TO * a_DT new_JJ x_NN to_TO * a_DT complete_JJ x_NN to_TO * new_JJ p_NNP to_TO * complete_JJ p_NNP a_DT new_JJ x_NN has_VHZ been_VBN a_DT complete_JJ x_NN has_VHZ been_VBN

Table 1: Patterns

hicle. Such division of the data set poses an interesting question whether these distinctions can be adequately captured by a method we employ. For extracting hyperonymy relation, there seems to be a consensus that very general information (like *John is a human*) is not likely to be found in the data. It is therefore unclear whether the 2-way clustering (*natural* vs. *artifact*) would provide accurate results. We hypothesize that a more granular distinction can be captured much better.

To conduct all experiments, we use the Web data, particularly, Google API to extract snippets. Similarly to the experiments by Cimiano and Wenderoth(2007), a number of extractions is set to 50. However, if enumerations and conjunctions are treated, a number of extractions per seed might be greater than this threshold. All snippets are tokenized and tagged by a PoS analyzer, which in our case is TreeTagger[2]. Further, the preprocessed data is matched against a given pattern. PoS information allows us to reduce a number of candidates for the qualia roles. Unlike Cimiano, we do not employ any ranking of the extracted elements but use them to build a word-space model. In such a model, rows correspond to the words provided by the organizers of the challenge and columns are the qualia elements for a selected role. As in most word-space models, the elements of a matrix contain frequency counts. CLUTO (Zhao and Karypic, 2002) toolkit is used to cluster the seeds given the information in matrices.

Table 2 presents the results (when *formal* role only is used) in terms of purity and entropy. Entropy of cluster i is usually measured as

$$e_i = -\sum_{j=1}^{m} p_{ij} log(p_{ij}) \qquad (1)$$

where m is a number of classes and p_{ij} is probability that an object in cluster i belongs to cluster j, $p_{ij} = n_{ij}/n_i$. Purity r is defined as $r = \max_j p_{ij}$. The total entropy (purity) is obtained by a weighted sum of the individual cluster entropies (purities).

[2]available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

clustering	entropy	purity
2-way	0.59	0.80
3-way	**0.00**	**1.00**
6-way	0.13	0.89
2-way$_{>1}$	0.70	0.77
3-way$_{>1}$	0.14	0.96
6-way$_{>1}$	0.23	0.82

Table 2: Performance using *formal* role only

The 2-way clustering resulted in the imperfect discrimination between natural and artifact categories. Errors are caused by classifying vegetables as artifacts while they belong to the category *natural kind*. The 3-way clustering was intended to exhibit differences among fruit and vegetables (1st category), birds and animals (2nd category) and tools and vehicles (3rd category). In contrast to the 2-way clustering, there have been no errors observed in the clustering solution. While conducting experiments with the 6-way clustering aiming at finding 6 clusters corresponding to the abovementioned categories, we have noticed that vehicles and tools are not properly discriminated.

We have not filtered the acquired formal role elements in any way. As there is noise in the data, we decided to conduct an additional experiment by removing all features with the frequency 1 (2-way$_{>1}$, 3-way$_{>1}$, 6-way$_{>1}$). We observe lower purity for all three categorization tasks which suggest that some of the removed elements were important. In general, seeds in such categories as *bird* or *animal* get many qualia elements with the high frequency for the *formal* role varying from very general such as *creature*, *mammal*, *species* to quite specific (*pheasant*, *vertebrate*). Some members of other categories such as *tool* or *vehicle* do not possess as many features and their frequency is low.

6 Discussion

To evaluate which features were important for each particular solution and shed light on problematic areas, we carried out some additional analysis. Table 3, Table 4 and Table 6 present descriptive and discriminative features for the $2-$way, $3-$way and $6-$ way clustering respectively. Descriptive features correspond to the features that describe a given cluster the best and the discriminative are those which highlight the differences between a given cluster and the rest. Each feature is provided with the percentage of its contribution to a given category. In each table A stands for a descriptive part and B denotes discriminative one.

The categories are not necessarily homogeneous and this can be observed given the descriptive features. For instance, the category *vegetables* includes the member *mushroom* the features of which are quite distinct from the features of the other members (such as *potato* or *onion*). This is reflected by the feature *fungi* in the descriptive part of *vegetables*. The category *bird* includes the false descriptive feature *story*. As mentioned, we did not rank extracted terms in any way and such titles as *owls and other stories* heavily contributed to the feature *story*. It turned out that such titles are frequent for the *bird* category and, fortunately, a presence of this feature did not result in any misclassifications.

Telic role Having hoped that additional information might be helpful for such categories as *tool* and *vehicle*, we added terms extracted for other qualia roles. Unfortunately, none of them drastically changed the overall performance. The fact that *telic* role does not have a considerable impact on the final performance can be explained if one looks at the extracted terms. As some examples in Table 5 suggest, patterns for the *telic* role provide useful cues to what a purpose of the given entity is. Nevertheless, accurate extractions are usually not shared by other members of the same category. For instance, various tools have a quite different purpose (e.g., *to cut*, *to hit*, *to serve*) and there is no common denominator which would serve as a descriptive feature for the entire category.

Constitutive role In contrast to the telic role, constitutive information is either too general or too specific and does not contribute to the overall performance either. For instance, it is known that mushrooms consist of mycelium, water and have a cap; telephones have transceivers and handsets; boats consist of cabins, engines, niches and hulls but this information is not really used to discriminate among categories.

Agentive role The last role to consider is an agen-

	Cluster	Features
A	NATURAL	animal (43.3%), bird (23.0%), story (6.6%), pet (3.5%), waterfowl (2.4%)
	ARTIFACT	tool (19.7%), fruit (14.6%), vegetables (10.0%), vehicle (9.7%), crop (5.1%)
B		animal (22.1%), bird (11.7%), tool (10.1%), fruit (7.4%), vegetables (5.1%)

Table 3: 2−way clustering: descriptive vs. discriminative features

	Cluster	Features
A	VEGETABLE	fruit (41.3%), vegetables (28.3%), crop (14.6%), food (3.4%), plant (2.5%)
	ANIMAL	animal (43.3%), bird (23.0%), story (6.6%), pet (3.5%), waterfowl (2.4%)
	ARTIFACT	tool (31.0%), vehicle (15.3%), weapon (5.4%), instrument (4.4%), container (3.9%)
B	VEGETABLE	fruit (21.0%), vegetables (14.3%), animal (11.6%), crop (7.4%), tool (2.5%)
	ANIMAL	animal (22.1%), bird (11.7%), tool (10.1%), fruit (7.4%), vegetables (5.1%)
	ARTIFACT	tool (15.8%), animal (14.8%), bird (7.9%), vehicle (7.8%), fruit (6.8%)

Table 4: 3−way clustering: descriptive vs. discriminative features

seed	extractions
helicopter	to rescue
rocket	to propel
chisel	to cut, to chop, to clean
hammer	to hit
kettle	to boil, to prepare
bowl	to serve
pencil	to draw, to create
spoon	to serve
bottle	to store, to pack

Table 5: Some extractions for the *telic* role

tive one. As pointed out by Pustejovsky (2001), it might be helpful to distinguish between natural kinds and artifacts. Indeed, by adding results delivered by agentive patterns to those corresponding to the formal role, entropy decreases to **0.09** and purity increases to **0.91** on 6-way clustering. However, the results do not change for the 2-way clustering still classifying the members of the category *vegetables* as artifacts. An interesting observation is that such division reflects a degree of animacy. It is well known that personal pronouns have the highest animacy followed by humans, animals, plants, concrete things and abstract things (in this order). Clustering based on the formal role alone or any combination of other roles with it always distinguishes well animals (a *groundAnimal* and a *bird*) from plants but it seems to cut the animacy hierarchy right after animals by placing vegetables to artifacts.

A combination of the agentive and the formal roles (Figure 1) finally correctly classifies *rocket* as a vehicle. We can also observe that agentive features (*to develop*, *to invent*, etc.) mostly influence such categories as vehicles and tools.

All features (rows) presented in Figure 1 were selected on the basis of the union of the discriminative and descriptive features per cluster. Actual seeds (words) are given in columns.

6.1 Error Analysis

We also analyzed the clusters by means of features that frequently occur together. In CLUTO, there are two possibilities to perform such analysis, either by employing cliques or itemsets. Our aim in this case is to find possible subclusters withing resulting clusters. If there are any subclusters, they can be either attributed to the existing subcategories of a given category or by clustering errors so that two subclusters were merged even though they do not belong to the same category. Such information would be especially useful for the *tool* and *vehicle* classes. A cluster containing a mixture of vehicles and tools is indeed described by three frequent itemsets, *container-load*, *container-appliances* and *vehicle-aircraft*. Consequently, such words as *bowl*, *bottle* and *kettle* are all correctly clustered together

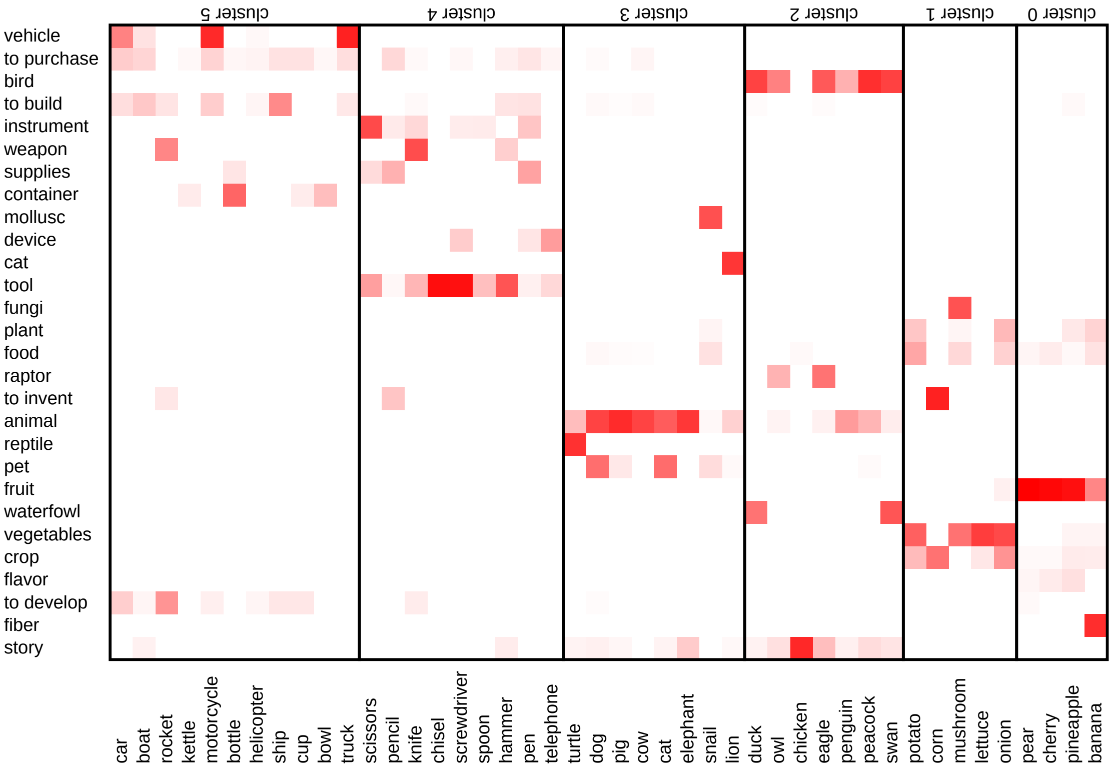

Figure 1: A combination of the formal and the agentive roles

but as a subcluster they occur in a wrong cluster which describes vehicles rather than tools.

CLUTO also provides a possibility to analyze how similar a given element is to other members in the same cluster. We expect a misclassified instance to be an outlier in a cluster. For this reason, we look at the z-scores of *rocket*, *bowl*, *cup*, *bottle* and *kettle*. There are two types of z-score, internal and external. Given an object j which belongs to the cluster l, the internal z-score is computed as follows:

$$z_I = \frac{s_j^I - \mu_l^I}{\delta_l^I} \tag{2}$$

In Eq. 2 s_j^I stands for the average similarity between the object j and the rest objects in the same cluster, μ_l^I is the average of s_j^I values over all objects in the lth cluster, and δ_l^I is the standard deviation of the similarities.

The external z-score is defined in a similar way with the only distinction that similarity is measured not with the object in the same cluster but in all other clusters.

The core of the cluster respresenting tools is formed by *chisel* followed by *knife* and *scissors* as they have the largest internal z-score. When the formal role only is used, the same cluster wrongly contains *rocket* but according to the internal z-score, it is an outlier (with the lowest z-score in the cluster). What concerns the ”container” subcluster is the cluster of vehicles, *bowl*, *cup*, *bottle* and *kettle* all have the lowest internal z-scores. The core of the cluster of vehicles is a *truck* and *motorcycle*.

By examining features we found several types of errors:

1. Errors occurring at the extraction stage

2. Lexical ambiguity and erroneous statements in text

The first type of errors can be further classified as errors due to the incorrect tagging or due to the imperfect extraction mechanism. In general, seeds we have can be PoS ambiguous and to take this ambiguity into account, we put strict restrictions on the pattern matching. None of the seeds which were tagged

	Cluster	Features
A	fruitTree	fruit (90.1%), fiber (5.2%), plant (0.6%), food (0.5%), flavor (0.5%)
	green	vegetables (56.4%), crop (25.2%), food (4%), fungi (3.6%), greens (2.8%)
	bird	bird (60.9%), story (10.7%), waterfowl (6.4%), animal (3.2%), raptor (3.1%)
	groundAnimal	animal (68.5%), pet (7.8%), reptile (2.5%), cat (2.4%), mollusc (1.8%)
	tool	tool (53.8%), weapon (9.4%), instrument (7.6%), supplies (5.4%), device (1.8%)
	vehicle	vehicle (42.8%), container (10.8%), aircraft (5.1%), appliances (4.9%), load (3.5%)
B	fruitTree	fruit (46.1%), animal (10.5%), tool (6.5%), bird (5.6%), vehicle (3.2%)
	green	vegetables (28.9%), crop (12.2%), animal (10.7%), tool (6.6%), bird (5.7%)
	bird	bird (34.7%), tool (8.4%), fruit (6.2%), vegetables (4.2%), vehicle (4.1%)
	groundAnimal	animal (31.4%), tool (8.7%), bird (7.4%), fruit (6.4%), vegetables (4.4%)
	tool	tool (28.1%), animal (12.4%), bird (6.6%), fruit (5.7%), weapon (4.9%)
	vehicle	vehicle (22.6%), animal (11.4%), tool (7.0%), bird (6.0%), container (5.7%)

Table 6: $6-$way clustering: descriptive vs. discriminative features

as verbs are matched in text which narrows down a number of extracted terms. However, our analysis reveals that in most cases seeds tagged as verbs are not ambiguous and these are the PoS tagging errors. Besides, we mostly rely on the PoS information and some errors occur because of the imperfect extraction. For instance, if PP attachment is not handled, the extracted terms are most often incorrect.

The second type of errors is attributed to the lexical ambiguity such as the sentence below:

(3) in fact, scottish gardens are starting to see many more butterflies including peacocks, . . .

From here, by using patterns for the formal role we get that *peacock* is a kind of *butterfly* which is not correct (according to the gold standard). However, we hope that such rare incorrect extractions will not play a significant role while clustering, if enough evidence is given. In our experiments *peacock* is always correctly categorized as a bird because of many other features which have been extracted. In particular, a feature *bird* has the highest frequency for this particular seed and, consequently, other rare features contribute to a lesser degree.

7 Conclusions

We presented a method for the concrete noun categorization which uses a notion of qualia structures. Our initial hypothesis of the difficulty of distinguishing between very general levels of categorization was supported by the empirical findings. While all tools and vehicles are always correctly identified as artifacts, vegetables are not classified as a natural category. The 3-way categorization provides the best performance by correctly identifying categories for all words. The 6-way categorization reveals difficulties in discriminating tools and vehicles. In particular, all containers are grouped together but incorrectly placed in the cluster *vehicle*. The most difficult element to classify is *rocket* which according to the gold standard is a vehicle. However, most features describing it are related to weapon and it is less surprising to find it in a category *tool* with such words as *knife* sharing this feature. When agentive role information is added, *rocket* is finally correctly classified as a vehicle.

Regarding qualia roles, the *formal* role is already sufficient to discriminate well between 3 categories. Adding extra information on other roles such as telic and constitutive does not improve results. By inspecting features which are extracted for these roles we can conclude that many of them are relevant (especially for the telic role) but there are hardly any members of the same category which would share them. This finding is in line with (Cimiano and Wenderoth, 2007) who mentioned the telic role as the one humans mostly disagree on.

As possible future directions, it would be interesting to experiment with other corpora. We have conducted all experiments with the Web data because of our intention to capture different qualia elements for each role. We noticed however that some patterns proposed by (Cimiano and Wenderoth, 2007) are more suitable for the artifacts than natural kinds. More general patterns such as those by (Yamada et al., 2007) might be more helpful for the telic role but all candidates must be ranked to select the best suitable.

References

Philipp Cimiano and Johanna Wenderoth. 2007. Automatic Acquisition of Ranked Qualia Structures from the Web. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007, pages 888-895*.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text data. *In Proceedings of COLING-92, pp. 539-545*.

Alessandro Lenci. 2000. Building an Ontology for the Lexicon: Semantic Types and Word Meaning. *In Ontology-Based Interpretation of Noun Phrases, Department of Business Communication and Information Science, University of Southern Denmark, Kolding:103-120* .

James Pustejovsky. 2001. Type Construction and the Logic of Concepts. *In The Syntax of Word Meaning, P. Bouillon and F.Busa (eds.) 2001, Cambridge University Press*.

James Pustejovsky. 2000. Syntagmatic Processes. *In Handbook of Lexicology and Lexicography, de Gruyter, 2000*.

Magnus Sahlgren. 2006. The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *PhD thesis, SICS Dissertation Series 44*.

Ichiro Yamada, Timothy Baldwin, Hideki Sumiyoshi, Masahiro Shibata and Nobuyuki Yagi. 2007. Automatic Acquisition of Qualia Structure from Corpus Data. *IEICE Transactions Inf. & Syst., vol. E90-D*.

ClaudiaZavaglia and Jualiana Galvani Greghi. 2003. Homonymy in Natural Language Processes: A Representation Using Pustejovsky's Qualia Structure and Ontological Information. *In Computational Processing of the Portuguese Language*.

Y. Zhao and G. Karypis. 2006. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. *In CIKM 2002*.