Query Interpretation – an Application of Semiotics in Image Retrieval

Maaike H.T. de Boer[1,2], Paul Brandt[1,3], Maya Sappelli[1,2], Laura M. Daniele[1], Klamer Schutte[1], Wessel Kraaij[1,2]

1: TNO
Anna van Buerenplein 1,
2595 DA The Hague,
The Netherlands
{maaike.deboer, paul.brandt,
maya.sappelli, laura.daniele,
klamer.schutte,wessel.kraaij}@tno.nl

2: Radboud University
Toernooiveld 200
6525 EC Nijmegen
The Netherlands
{m.deboer, w.kraaij,
m.sappelli}@cs.ru.nl

3: Eindhoven University of Technology
De Groene Loper 19
Eindhoven
The Netherlands
p.brandt@tue.nl

Abstract— **One of the challenges in the field of content-based image retrieval is to bridge the semantic gap that exists between the information extracted from visual data using classifiers, and the interpretation of this data made by the end users. The semantic gap is a cascade of 1) the transformation of image pixels into labelled objects and 2) the semantic distance between the label used to name the classifier and that what it refers to for the end-user. In this paper, we focus on the second part and specifically on (semantically) scalable solutions that are independent from domain-specific vocabularies. To this end, we propose a generic semantic reasoning approach that applies semiotics in its query interpretation. Semiotics is about how humans interpret signs, and we use its text analysis structures to guide the query expansion that we apply. We evaluated our approach using a general-purpose image search engine. In our experiments, we compared several semiotic structures to determine to what extent semiotic structures contribute to the semantic interpretation of user queries. From the results of the experiments we conclude that semiotic structures can contribute to a significantly higher semantic interpretation of user queries and significantly higher image retrieval performance, measured in quality and effectiveness and compared to a baseline with only synonym expansions.**

Keywords– query expansion; natural language queries; image retrieval; semantic reasoning; computational semiotics.

I. INTRODUCTION

More and more sensors connected through the Internet are becoming essential to give us support in our daily life. In such a global sensor environment, it is important to provide smart access to sensor data, enabling users to search semantically in this data in a meaningful and, at the same time, easy and intuitive manner.

Towards this aim, we developed a search engine that combines content based image retrieval (CBIR), Human Media Interaction and Semantic Modelling techniques in one single application: "Google® for sensors" or "GOOSE" for short. This paper builds on our earlier work on applying semantic reasoning in image retrieval [1] in the GOOSE search engine, an overview paper of which is given in [2][3].

A major issue to text searches in visual data is "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation", coined by [4] as the semantic gap in CBIR. Our application is able to retrieve visual data from multiple and heterogeneous sources and sensors, and responds to the fact that the semantic gap consists of two parts [5]: the first part addressing the realm where raw image pixels are transformed into generic objects to which labels are applied to represent their content; the second part addressing the realm of semantic heterogeneity, representing the semantic distance between the object labelling and the formulation by the end-user of a query that is meant to carve out a part in reality that situates that object. The GOOSE approach to closing the first section addresses image classification and quick image concept learning, presented in [6], and fast re-ranking of visual search results, presented in [7]. This paper addresses the second section of the semantic gap by applying, in this order, query parsing, concept expansion, and concept mapping to labels associated with certain classifiers. Query concepts that do not match any classifier's label are expanded using an external knowledge base, in this case ConceptNet [8], to find alternative concepts that are semantically similar to the original query concepts but do match with a classifier label.

Whereas in [1] we only used 'IsA' and 'Causes' relations in the query expansion, in this work we address additional types of relations in order to improve the matching rate between query concepts and classifier labels. However, a drawback of considering additional relations is that the algorithms for their semantic interpretation become tightly coupled to the particular external knowledge base of choice, rendering them less applicable for other knowledge bases. To overcome this limitation and keep our semantic interpretation generically applicable to other knowledge bases, we introduce the use of semiotics that provides guidance to how humans interpret signs and how the abstract relationships between them apply. Due to its universal application, a semiotic approach not only provides us with the flexibility to use different knowledge bases than ConceptNet, but it is also independent from domain-specific terminologies, vocabularies and reasoning. By defining a simple mapping from the specific relationships of the knowledge base of choice, e.g., ConceptNet, onto semiotic structures, the semantic interpretation algorithms can latch onto the semiotic structures only. The resulting transparency between, at the one hand, the semantic interpretation algorithms and, at the other hand, abstracting from the

specifics of (i) the relationships that are available in the external knowledge base, and (ii) the domain-specific vocabularies, brings about the required general applicability of our solution.

Summarizing, we seek to improve the matching rate between query concepts and classifier labels, by 1) considering more, if not all, relations that are available in a knowledge base; while remaining 2) as independent to the external knowledgebase as possible; and 3) as computationally lean as possible.

We formulate our research question as

To what extent can semiotic structures contribute to the semantic interpretation of user queries?

In order to answer our research question, we conducted an experiment on our TOSO dataset [9], which contains 145 test images and 51 trained classifiers. Furthermore, for evaluation purposes, we defined 100 user queries [10]. We annotated these user queries with their ground truth for both parts of the semantic gap: (i) the ground truth for semantic matching, identifying the classifier labels that are meant to be found for each user query, and (ii) the ground truth for the image retrieval, identifying the images that are meant to be found. For the different types of semiotic structures we calculated the effectiveness and quality in terms of different types of F-measure for both semantic matching and image retrieval.

From the results of these experiments, we can conclude that applying semiotic relations in query expansion over an external, generic knowledge base contributes to a high quality match between query concepts and classifier labels. It also significantly improves image retrieval performance compared to a baseline with only synonym expansions. Some relations that are present in ConceptNet could not be assigned to the applied semiotic structures; inclusion of these relations in the semantic analysis provided for higher effectiveness at the cost of losing loose coupling between these relations and the algorithms that implement the semantic analysis. However, we did not investigate other potential semiotic structures to this effect.

The main contribution of this paper is a generic approach to the expansion of user queries using general-purpose knowledge bases, and how semiotics can guide this expansion independently from the specific knowledge base being used.

This paper is structured as follows: Section II describes related work on query expansion and semiotics; Section III presents a short tutorial on semiotics; Section IV provides an overview of the generic semantic interpretation system; Section V explains the semantic analysis and how we have positioned semiotic structures for its support; Section VI describes the experiment that has been performed with the application, followed by a presentation and discussion of their results in Sections VII and VIII, respectively. We conclude our work, including indications for future work, in Section IX.

II. RELATED WORK

In this section, we discuss related work in CBIR about the first part of the semantic gap, i.e., automatic classifier annotation, as well as the second part of the semantic gap, i.e., some efforts related to query expansion using semantic relations. Finally, we discuss related work on computational semiotics.

A. Automatic image annotation

Most of the effort in applying semantics in CBIR is aimed at training classifiers using large sources of visual knowledge, such as ImageNet [11] and Visipedia [12]. The trained classifiers are subsequently annotated with one or more labels that should describe their meaning. However, these annotations are often subjective, e.g., influenced by the domain of application and not accurate from a semantic point of view. Consequently, users that apply these classifiers need to have prior knowledge about the context of use of the annotations. In order to overcome this issue and facilitate the use of classifiers without the need of training, various efforts in the literature focus on improving the annotations. These efforts mainly apply domain-specific ontologies as basis for annotation, such as the ontologies in [12][13] that are used to annotate soccer games, or for the purpose of action recognition in a video surveillance scenario [15]. Although these approaches provide for more intuitive semantics that require less prior knowledge from the user, they are tailored to specific domains and cannot be reused for general-purpose applications.

B. Relation-based query expansion

Several systems proposed in the literature address query expansion exploiting relations with terms that are semantically similar to the concepts in the query [16][17][18]. The system in [16] facilitates natural language querying of video archive databases. The query processing is realized using a link parser [19] based on a light-parsing algorithm that builds relations between pairs of concepts, rather than constructing constituents in a tree-like hierarchy. This is sufficient for the specific kind of concept groups considered in the system [16], but is limitative for more complex queries.

The Never Ending Image Learner (NEIL) proposed in [17] is a massive visual knowledge base fed by a crawler that runs 24 hour a day to extract semantic content from images on the Web in terms of *objects*, *scenes*, *attributes* and their *relations*. The longer NEIL runs, the more relations between concepts detected in the images it learns. Analogously to our approach, NEIL is a general-purpose system and is based on learning new concepts and relations that are then used to augment the knowledge of the system. Although NEIL considers an interesting set of semantic relations, such as taxonomy (*IsA*), partonomy (*Wheel is part of Car)*, attribute associations (*Round_shape is attribute of Apple* and *Sheep is White)*, and location relations (*Bus is found in Bus_depot*), most of the relations learned so far are of the basic type 'IsA' or 'LooksSimilarTo'.

Furthermore, in [18] knowledge bases ConceptNet and Wikipedia, and an expert knowledge base are compared for

semantic matching in the context of multimedia event detection. Results show that query expansion can improve performance in multimedia event detection, and that the expert knowledge base is the most suitable for this purpose. When comparing Wikipedia and ConceptNet, ConceptNet performs slightly better than Wikipedia in this field. In their comparison, the authors only considered query expansion using the ConceptNet 'IsA' relation.

C. Semiotics in CBIR

Although text analysis is its primary field of application, recently semiotics gained the interest in the field of ICT. The application of semiotics in computer science is best illustrated with the emergence of computational semiotics, where a clear starting point for its definition is the fact that signs and sign systems are central to computing: manipulation of symbols applies to everything that happens in computer science, from user interfaces [20][21][22] to software engineering [23][24][25][26], from model-driven engineering [27] to conceptual and knowledge modelling [28][29][30], and interoperability [31] alike. In relation to CBIR, many studies, summarized by [5], accept the existence of 'semantic layers' in images. Every layer provides for another abstraction and aggregation of the things that are being denoted. The studies referenced in [5] address these layers as distinct realms, and act accordingly by constraining themselves to one layer. However, semioticians address these layers as a whole, and study it as a process to which they refer as *unlimited semiosis* (see next section). We are inspired by that approach and therefore part of our work considers unlimited semiosis as algorithmic foundation when addressing these layers. Application of semiotics in CBIR and especially about user query interpretation is very limited, and the following two studies represent, to the best of our knowledge, good examples of its main focus.

Yoon [32] has investigated the association between denotative (literal, definitional) and connotative (societal, cultural) sense-making of image meta-data in support of image retrieval. This approach is similar to ours in that it is based on semiotics structures to bridge the semantic gap. Although the results are promising, it cannot be applied in our generic context due to the domain-specific foundations that are implicit to connotations.

Closely related to it, [33] studies how semiotics can account for image features that characterize an audio, visual or audio-visual object, in order to facilitate visual content description or annotation. Their model integrates low-level image features such as color and texture together with high-level denotative and connotative descriptions. This approach differs with ours in that they do not make a distinction between the two cascading parts of the semantic gap, but instead take an integrated approach.

III. SEMIOTICS

We include a brief tutorial on semiotics here since we believe this discipline is not very well known to our readers. Specifically, we address the semiotic structures that we apply. According to semiotics, humans make meanings through our creation and interpretation of signs [34]. A sign can be anything, varying from a character to a sculpture, as long someone interprets it, i.e., it goes beyond the sign itself. A *semiotic sign*, or sign for short, represents a structure. In Peirce's semiotic triangle [35] it consists of three closely related aspects, as depicted in Figure 1(a).

The *Representamen* (sometimes denoted as the *sign vehicle*) represents the form that the sign takes, e.g., this paper or a Chinese character. This form can be written, spoken or displayed, such as a picture or movie scene. The *Interpretant* in Figure 1(a) does not refer to an interpreter but rather to the sense given to the sign, i.e., our mental representation of reality, such as the mental "picture" of a 'red apple' that one has in mind. The *Object* in Figure 1(a), is the concrete thing in reality to which the sign refers, where this reality may also be an hypothetical reality, e.g., a unicorn. As opposed to the direct relationships between the interpretant and the object, and the interpretant and the representamen, which are drawn as solid lines, the relationship between the object and representamen is not direct, and hence depicted with dots. A semiotic sign only qualifies as such when it unifies all three aspects into a meaningful ensemble: the object is perceived by our senses and abstracted into the interpretant, which subsequently is represented by the representamen. A Peircean sign concurrently indicates what is being represented, how it is being represented and how it is being interpreted. The sense making is subjective by nature, hence every actor makes use of their own signs, although the sign's representamen can be shared.

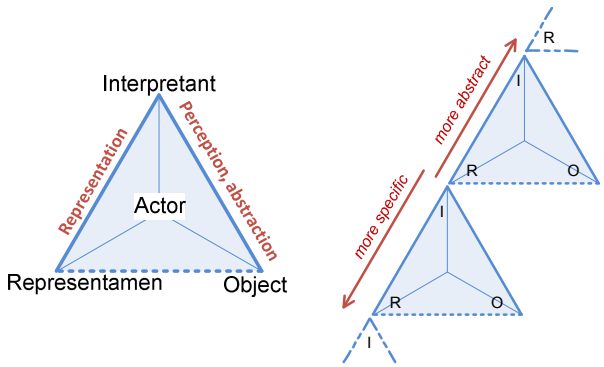


Figure 1. Peirce's semiotic triangle (a), unlimited semiosis (b).

Although semiotic techniques are used mainly to analyze texts, in this paper we extend their usage to the semantic analysis of user queries. Specifically, we consider the semiotic structures that help us to better distinguish between the relations defined in external knowledge bases, e.g., to which extent is an 'IsA' relation semiotically different from a 'Causes' relation, and how can we can benefit from this difference. We selected the semiotic structures *unlimited semiosis*, *paradigms* and *syntagms* from [34] as vehicles to a universal approach towards reasoning over different semantic relationships defined in various knowledge bases.

A. Unlimited semiosis

Sense making is above all a process, and Peirce refers to the interaction between the three elements of the semiotic triangle as 'semeiosis' [21][22]. He also observes that semiotic signs are coupled: "*a sign (...) creates in the mind of that person an equivalent sign, or perhaps a more developed sign.*" (ibid.). Consequently, the interpretant at level N is yet another representamen but at a 'more developed' level N+1. Eco [36] uses the term 'unlimited semiosis' to refer to the succession of cascading signs that emerge from that, ad infinitum (Figure 1(b)). The application of unlimited semiosis gives us the capability to address semantic issues that relate to the different levels of *granularity* between the query of concept and a classifier label, e.g., 'vehicle' (higher level of detail) and 'car' (lower level of detail).

B. Paradigms

From a semiotic perspective, semantics arise from the differences between signs. In other words, without their ability to signify differences, signs could not carry meaning at all. Differences of signs concern two distinctions, as depicted in Figure 2. The first distinction, called *paradigms*, concern substitution and signify functional contrasts, e.g., how to differentiate the sentences 'the man cried' from 'the woman cried'. Signs are in paradigmatic relation when the choice of one ('man') excludes the choice of another ('woman') [37], i.e., disjunction. The selection of a particular sign from a paradigmatic set, e.g., selecting *man* from *{man, woman, child}*, implies an intentional exclusion of the interpretations that originate from the use of the other signs from the paradigmatic set, e.g, *woman* and *child*. Other paradigms in Figure 2 are {*cry, sing, mutter*}, and one about cardinality. Due to their nature, paradigms provide us with the ability to reject alternative concepts.

Figure 2. The semantics of a sign is determined by both its paradigmatic and syntagmatic relations.

C. Syntagms

Where paradigms reflect differences concerning substitution, the second axis in Figure 2, *syntagms*, reflect differences concerning position, e.g., the position of words in a sentence, or the position of paragraphs in a section. This reflects how the juxtaposition of the distinct parts (the signs) complete into a whole, e.g., how words take their place in a grammatically correct sentence, or how chapters are used to form a book. Syntagmatic relations reflect the admissible combinations of paradigmatic sets into well-formed structures, e.g., conjunction. The use of one syntagmatic structure over another influences semantics, e.g., 'the ship that banked' versus 'the bank that shipped' use identical signs, whilst their positions in the sentence turn them from a verb to an object and vice versa, with completely different semantics as result.

IV. Generic Semantic Reasoning System

Figure 3 shows an overview of the semantic reasoning parts of the GOOSE system in which green and blue parts represent the components that realize the semantic reasoning, yellow parts represent the components dedicated to the image classification task and the white parts represent external components. The image classification task, which is elaborated in [6], captures the semantics of visual data by translating the pixels from an image into a content description (which could be a single term), coined as *annotated images*.

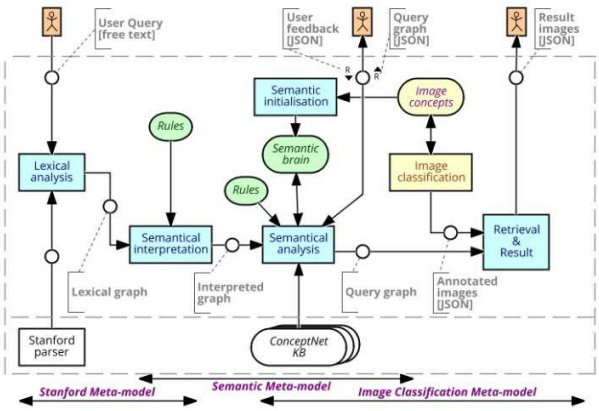


Figure 3. System overview.

The semantic reasoning starts with a user query in natural language. The query is processed by four modules, while a fifth module takes care of initializing the system and learning new concepts. In the first stage, the query is sent to the *Lexical Analysis* module that parses it using the Stanford Parser [38]. The Stanford Parser returns a lexical graph, which is used as input to the *Semantic Interpretation* module. In this module, a set of rules is used to transform the lexical elements of the Stanford meta-model into semantic elements of an intermediary ontology (our meta-model, discussed in section IV.C below). The interpreted graph is sent to the *Semantic Analysis* module that matches the graph nodes against the available image concepts. If there is no exact match, the query is expanded using an external knowledge base, i.e., ConceptNet, to find a close match. The interpretation resulting from the Semantic Analysis is presented as a query graph to the user. The query graph is also used as input for the *Retrieval and Result* module, which provides the final result to the user. In the following subsections the complete process is described in detail using the sample query *find a red bus below a brown animal*. In this particular query, its positional part, e.g., *below*, should be understood from the viewpoint of the user posing the

query, i.e., the relative positions of the 'red bus' and the 'brown animal' as shown at the user's screen.

A. Semantic Initialization

This module provides an initial semantic capability by populating the Semantic Brain, which holds all *image concepts* that are known to the system. Image concepts are represented as instances of the meta-model (discussed in Section IV.C), and refer to those things that the image classification task is capable of detecting. This component also handles updates to the Semantic Brain following from new or modified image classification capabilities and semantic concepts.

B. Lexical Analysis

In the Lexical Analysis module, the user query is lexically analyzed using the Typed Dependency parser (englishPCFG) of Stanford University [38]. Before parsing the query, all tokens in the query are converted to lower case. In the example of *find a red bus below a brown animal*, the resulting directed graph from the Lexical Analysis is shown in Figure 4.

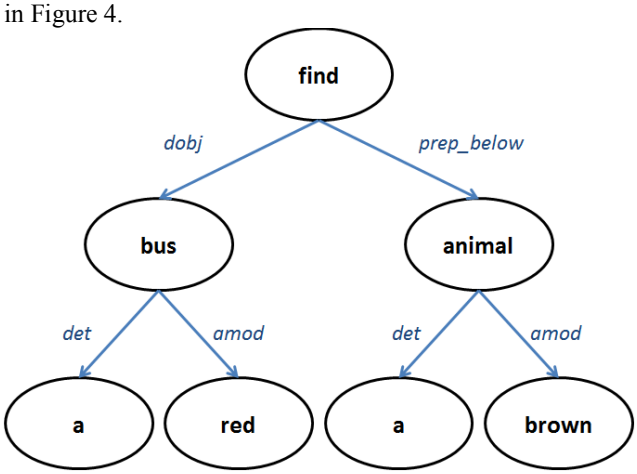


Figure 4. Lexical Graph.

C. Semantic Interpretation

Since GOOSE is positioned as a generic platform, its semantics should not depend on, or be optimized for, the specifics of one single domain of application. Instead, we apply a generic ontological commitment by defining a semantic meta-model, shown in Figure 5, which distinguishes objects that might (i) bear attributes (*a yellow car*), (ii) take part in actions (*a moving car*), (iii) occur in a scene (*outside*), and (iv) have relations with other objects, in particular ontological relations (*a vehicle subsumes a car*), spatial relations (*an animal in front of a bus*), and temporal relations (*a bus halts after driving*).

In the Semantic Interpretation module, a set of rules is used to transform the elements from the lexical graph into *objects*, *attributes*, *actions*, *scenes* and *relations*, according to the semantic meta-model in Figure 5.

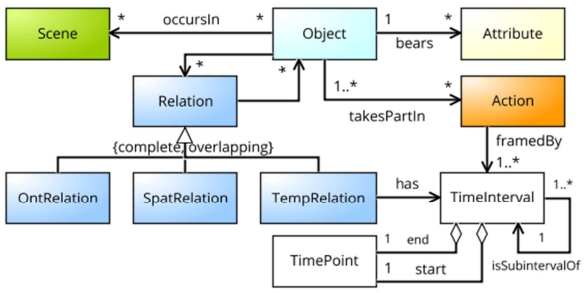


Figure 5. Semantic meta-model.

These rules include the following examples:

- Derive *cardinality* from a *determiner* (*det* in Figure 4), e.g., *the* in a noun in the singular form indicates a cardinality of 1, while *a/an* indicates at least 1;
- Derive *attributes* from *adjectival modifiers* (*amod* in Figure 4), i.e., adjectival phrases that modify the meaning of a noun;
- Derive *actions* from *nominal subjects* and *direct objects* (*nsubj* and *dobj*, absent in Figure 4), i.e., the subject and object of a verb, respectively;
- Actions that represent the query command, such as *find*, *is, show* and *have*, are replaced on top of the tree by the subject of the sentence.

The output of the Semantic Interpretation for the sample query *find a red bus below a brown animal* is shown in Figure 6.

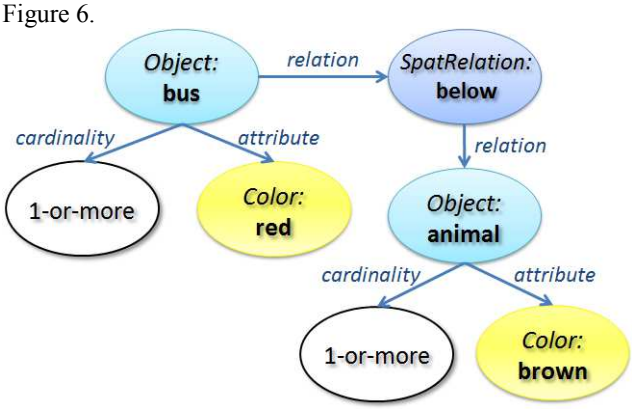


Figure 6. Interpreted Graph.

D. Semantic Analysis

The purpose of the Semantic Analysis is to align the elements from the interpreted graph, which are the query concepts, with the image concepts that are available in the Semantic Brain. For those objects, actions, scenes or attributes from the graph that do not have a syntactical identical counterpart ('exact match') in the Semantic Brain, and hence cannot be recognized by the image classification component, the query concepts are expanded into *alternative concepts* using an external general-purpose knowledge base. We use the external knowledge base ConceptNet to find these alternative concepts. Our principle of genericity and

loose coupling, however, facilitates the use of other or even more knowledge bases without the need to adapt the semantic analysis algorithms.

E. Retrieval and Result

This module retrieves the images that, according to the classifiers, contain concepts that carry an identical label as the query concepts (or alternative concepts). Furthermore, the cardinality, attribute and spatial relations should match with the query. If the image contains too many instances, the image is still included, however, with lower ranking. The spatial relations are determined by the edges of the bounding box. Because our bounding boxes are not accurate, we use a relaxed version of the prepositions. The upper left edge of the bounding box has the values [0,0]. In the preposition *left of*, the left edge of the bounding box of the right object should be right of the left edge of the bounding box of the left object, denoted as:

Left of	a.min.x < b.min.x
Right of	a.max.x > b.max.x
On top of	a.min.y < b.min.y
Below	a.max.y > b.max.y
And	a *and* b

V. Semiotic Structures in Semantic Matching

In this section, we explain how the semiotic structures introduced in Section II are used to implement the Semantic Analysis module of the GOOSE system.

We consider the external knowledge base, e.g., ConceptNet, to represent a graph. Query expansion is then similar to a graph traversal, where nodes represent alternative concepts, and edges represent relations between concepts. Each edge is of a specific type, e.g., *IsA*, *HasA*, *PartOf*, and more. Considering edges of particular types only, results in considering a subgraph. We apply a semiotic structure by only considering edges which type can be considered to represent that semiotic structure. This results in a subgraph, the characteristics of which corresponds to a large extend to the semiotic structure of choice. For example, by considering paradigmatic relations only, a paradigmatic subgraph emerges. Query expansion now becomes *semiotic subgraph* traversal, and will apply distinct traversal strategies that fit best the particular semiotic subgraphs. Additionally, we can also apply results from one subgraph traversal for traversals of other subgraphs, since their semiotic structures can be put into a specific relation to each other.

Whether an external knowledge base indeed will show these emerging semiotic subgraphs cannot be enforced, hence deviations should be anticipated. For instance, our algorithms will need to take into account the presence of loops in the subgraph, despite the fact that a semiotic theory might predict the emergence of a non-cyclic graph. The following sections will elaborate on the specifics for each semiotic structure.

A. Unlimited semiosis

This structure discerns relations that bear a direction-oriented application during query concept expansion, i.e., a relation expresses direction towards more *abstract* or more *specific* concepts. Using unlimited semiosis, we are able to select a certain expanded concept that is either more abstract, or more specific than the original query concept. This has consequences for the corresponding classifier. The selection of a more abstract alternative concept implies that the corresponding classifier will have less granularity and is therefore more general but less accurate from a semantical point of view. In other words, the *unlimited semiosis subgraph* that emerges, represents a directed, non-cyclic multigraph. For example, requesting for a 'flower' returns the 'plant' concept as alternative, which results in classifiers of all sorts of plants, including evergreens and bloomers alike. Vice-versa, classifiers that match a more specific concept have more granularity; are less general with a higher accuracy, e.g., 'flower' now returns 'rose' and the results will always be a flower albeit a specific type of flower, i.e., roses and not a bracken, but nor another type of flower, e.g., a tulip. Concept expansion, hence, can use each edge in either way, downstream or upstream. In effect, when searching for a more granular concept, it selects the *target node* of *more specific* edges (downstream), and selects the *source node* of *more abstract* edges as well (upstream). For less granular concepts it takes the converse approach, flipping the downstream and upstream directions. Furthermore, when traversing more abstract edges subsequently, semantics will expand gradually by including more and more other categories of concepts. Vice versa, when traversing more specific edges subsequently, semantics will reduce by excluding categories of concepts that do not concur with the added details. Therefore, edge traversals in the unlimited semiosis subgraph decreases the semantic correspondence with the original query concept and should be avoided as long as possible. This coincides with a breadth-first approach, in which one single iteration over the knowledge base will address all children and all parents as alternative concepts first, before considering grandchildren and grandparents in the next iteration.

In our implementation, we use the following relations from ConceptNet: *IsA*, *hasSubEvent*, *PartOf* and *HasA*. The 'IsA' and 'PartOf' relations are directed towards more abstraction, whilst the 'HasSubEvent' and 'HasA' relations are directed towards the more specific concepts.

B. Paradigms

As explained in Section III.B, paradigms represent sets of disjoint concepts. When considering only relations in the external knowledge base that express paradigms, we consider the subgraph that emerges as a non-directed graph. In this *paradigmatic graph* we consider the query concept to represent the one and only connecting node of otherwise disconnected (undirected) graphs, each of them representing a paradigm. In other words, the paradigms for the query concept are constructed by performing a depth-first approach, each single branch from the query concept leading to another paradigm.

Application of *paradigms* provides us with the ability to reject alternatives, because that is the nature of paradigms: the user made a conscious choice for this query concept and therefore specifically *excludes* the paradigmatic alternatives.

That implies that for every classifier label that has a match with an alternative concept, that classifier is considered to be a paradigm of the query concept and hence is excluded from the results. For an additional application of paradigms, consider the alternative concepts that result from graph traversals from one of the other methods. These alternative concepts are checked whether they are paradigmatic to the query concept. If so, not only that alternative concept is rejected but the whole branch that is accessed through that concept is pruned from the search space. In this way, paradigms are applied with the aim to reduce the combinatorial explosion that occurs in the graph traversals of other methods. We currently conduct research into this additional application of paradigms – due to time and space restrictions we do not present the results of this experiment in this paper. In our ConceptNet example we consider *MemberOf* and *DerivedFrom* as paradigmatic relations.

C. Syntagms

The application of syntagms is not restricted to the semantic analysis. For instance, the use of the Stanford parser to decompose the natural language query into a structure of related query concepts represents an example of the use of the *syntagmatic* structure, applying linguistic rules. Another example is the translation of the query concepts (and subsequently their alternatives) into an instantiation of the meta-model. Here, the relations that exist between the entities in the meta-model (Figure 5) provide for the allowed syntagmatic combinations.

Application of syntagms to the semantic analysis relates to their power to facilitate transitions between realms of classifiers, as follows. Because each classifier is bound to only one entity in the meta-model, e.g., objects, by application of syntagms we can search for alternative concepts that belong to other entities of the meta-model, e.g., actions, or properties. In this way, we enable an otherwise 'passive' set of classifiers for alternative concepts. An example of this is the expansion of the action 'person driving' to objects such as 'car', 'vehicle', or 'bike'. The knowledge that 'driving' relates to these objects is available in the knowledge base, and the syntagmatic relations reveal that knowledge. We conclude that the emerging *syntagmatic subgraph* is a directed, cyclic subgraph, in which edges represent transitions between entities in the meta-model.

In our ConceptNet example we consider the following relations to represent syntagms: *CapableOf*, *UsedFor*, *CreatedBy* reflect transitions from objects to actions; *Causes* reflect a transition from object to action; *hasProperty* from object to property.

VI. EXPERIMENT

In order to answer our research question *to what extent can semiotic structures contribute to the semantic interpretation of user queries?* we conducted an experiment.

In this experiment, we measure effectiveness and quality of different semiotic structures on the level of both semantic matching and image retrieval. The variable of the experiment is therefore represented by the differences in query expansion strategy, their core being the semiotic structures that are explained in the previous section. The experiment context is defined by our TOSO dataset and 100 manually defined queries. More information on the TOSO dataset can be found in subsection A. The type of queries can be found in subsection B. The experiment variations and its baseline are explained in subsection C. The design of the experiment if presented in subsection D, and its evaluation is explained in subsection E.

A. Dataset

The TOSO dataset [9] consists of 145 images of toys and office supplies placed on a table top. In these images multiple objects can be present in several orientations as well as objects of the same type with different colors. In Figure 7, a sample of the dataset has been depicted. Examples of these objects are different types of cars, a bus, an airplane, a boat, a bus stop, a traffic light, different types of traffic signs, Barbies with different colored dresses, different colored plants, a water bottle, a screwdriver, a hamburger and a helmet. For this dataset, 40 relevant object classifiers, trained on table top images, are available as well as 11 attribute classifiers, which are colors. The object classifiers are trained with a recurrent deep convolutional neural network that uses a second stage classifier [6]. The colors are extracted using [39].

Figure 7. A sample of the TOSO dataset.

B. Queries

In this experiment, we created 100 queries. In the definition of the queries we used our prior knowledge of the available classifiers by intentionally choosing interesting expansions, for example, their synonyms or hypernyms. This was done by searching online thesauri, independently from our ConceptNet example. In this way, we created a set of queries that does not have direct matches to the available classifiers, but for which the use of semiotic structures could be helpful. These queries are divided into five equal groups based on their semiotic or semantic structure as follows:

1) Synonym: synonyms of our labels;
   find the auto (classifier label: car);
2) Unlimited semiosis: hyponyms or hypernyms, i.e., parents or children of a label, or suspected 'part of' relations;
   find the Mercedes (classifier label: car)
   find the animal (classifier label: giraffe)
   find the leaf (classifier label: plant)
3) Paradigm: excluding brothers and sisters in the graph (man vs. woman), restrictions to objects by color and/or spatial relations;

find the air vehicle (as opposed to land vehicle, e.g., car, bus, tram);
find the red sign on the right of the yellow car;

4) Syntagm: actions and properties related to our labels;
find the things landing (classifier label: airplane);
find the expensive things (classifier labels: airplane, car);

5) Other: words which have a less clear or vague relation with a classifier label:
find the flower pot (classifier label: plant)
find the traffic jam (classifier label: cars)

For each of the queries, we established a semantic ground truth as well as an image ground truth. The semantic ground truth was established by manually annotating for each classifier label in our classifier set whether it is irrelevant (0) or relevant (1) to the query. In our annotation, a classifier is *relevant* if (i) a classifier label is syntactically similar to a concept in the query, or (ii) a classifier label represents a synonym of a query concept. For the image ground truth we used the 145 test images from the TOSO dataset. An external annotator established the ground truth by annotating, for each query and for each image, whether the image was irrelevant (0) or relevant (1) to the query. Establishing relevancy was left to the annotator's judgement. For both the semantic and image annotations, the instructions indicated that all cases of doubt should be annotated as relevant (1).

C. Experimental variable

In the experiment, we compare the following query expansion methods, the implementation of which has been explained in Section IV:

1) SYNONYM (baseline)
2) UNLIMITED SEMIOSIS
3) PARADIGM
4) SYNTAGM
5) ALL

In the first method, which represents our baseline, we use the basic expansion over specific relations that are found in ConceptNet: *Synonym* and *DefinedAs*. In methods 2 (UNLIMITED SEMIOSIS), 3 (PARADIGM) and 4 (SYNTAGM), we apply query expansion by traversing only the relations that are particular to the subject semiotic structure (defined in Section V), however we added the relations from the baseline. In the ALL method (5) all possible relations from ConceptNet, excluding *TranslationOf* and *Antonym,* are applied for query expansion.

D. Experiment design

The design of the experiment is based on the hypothesis that a query will be served best by a query expansion strategy that shares its semiotic structures, e.g., the SYNTAGM expansion method will find most mappings for syntagm queries and perform worse for other queries. Therefore, in our experiment each expansion method from Section V will apply its one single expansion strategy over all query groups from Section VI.B; different methods will therefore perform differently, i.e., result in different mapping counts.

In order to test our hypothesis, we designed and ran two evaluation cases. The first evaluation case addresses the part of the semantic gap that is about *semantic matching*. This case shows the impact of using semiotic structures on the effectiveness and quality of the mapping from the query to the classifier labels. The second evaluation case addresses the part of the semantic gap that is about *image retrieval*. This case shows the impact of semiotic structures on the effectiveness and quality of a full general-purpose image search engine.

E. Evaluation criteria

In our evaluations, we calculate the effectiveness and quality in terms of different types of F-measure for each query from Section VI.B. The following provides more detail for each evaluation case.

1) Semantic Matching In order to show the result of the expansion method on the mapping from the query to the classifier labels, we compare the result of each of the methods against the ground truth. This result is a list of classifier labels that are found by searching ConceptNet using the relations that are characteristic for the subject expansion method. In the evaluation we use two kind of metrics, corresponding to quality and effectiveness. The typical metric for quality is using precision, denoted P_{sg}, which takes into account the amount of true positives, i.e., found and annotated as relevant labels, and the total amount of found labels, i.e., true positives and false positives together, denoted as TP and FP, respectively:

$$P_{sg} = \frac{1}{n} * \sum_{q=1}^{n} \frac{TP_{sg}}{TP_{sg} + FP_{sg}}$$

where *n* denotes the total amount of queries.

The typical metric for measuring effectiveness is recall, denoted R_{sg}, which takes into account the amount of correctly found labels, i.e., true positives and the total amount of relevant labels, i.e., true positives and false negatives together, the latter denoted as *FN*:

$$R_{sg} = \frac{1}{n} * \sum_{q=1}^{n} \frac{TP_{sg}}{TP_{sg} + FN_{sg}}$$

where *n* denotes the total amount of queries.

Precision and recall are always an interplay, so we decided to not use precision and recall separately, but combine them by means of applying the F-measure. Since different applications can value the precision and recall of the semantic matching differently, the F_β-measure can be used to express that one should attach β times as much value to the recall results of the semantic matching than to its precision results. By using the F_β-measure as our primary means of evaluation, we can show the impact of the experiment results on three classes of applications, i.e., high quality applications that value precision over recall, high effectiveness applications that value recall over precision, and neutral applications that value precision equally important as recall. The F_β -measure is defined as:

$$F_\beta = (1 + \beta^2) * \frac{P_{sg} * R_{sg}}{(\beta^2 * P_{sg}) + R_{sg}}$$

For high quality applications, we put 10 times more emphasis on the precision and choose to use $\beta = 0.1$. For neutral applications we use the basic F-measure, i.e., $\beta = 1$ and for high effectiveness applications, we value recall 10 times more than precision and use $\beta = 10$. Naturally, these choices for β are made in order to show relative trends as opposed to an absolute judgement.

2) Image Retrieval The annotations are used in a similar way as on the level of the semantic matching. Again, F-score with $\beta = 0.1$ is used for high quality applications, $\beta = 1$ for neutral applications and $\beta = 10$ for high effectiveness applications.

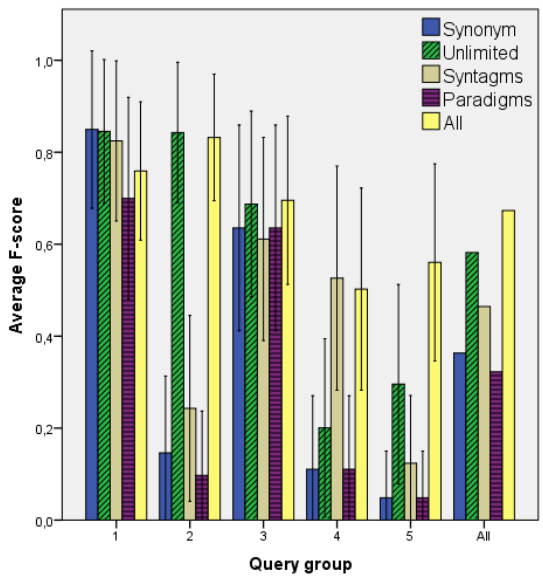
VII. RESULTS

In this section, we show the results of our experiment. The sections have the same structure as Section VI.D, so the first section explains the results about the semantic matching and the second section is about the results of the image retrieval.

For each of the evaluations, the assumption of normality was violated, as indicated by significant Kolmogorov-Smirnov statistics. We therefore present nonparametric Friedman-tests and Wilcoxon Signed-Ranks Tests to compare the different methods.

A. Semantic Matching

1) High precision system ($\beta = 0.1$): Graph 1 shows the F-score for the high precision system for each of the methods for each type of query group with the confidence interval of 95%. For two queries, both in group 4, no relevant annotation was available, so in group 4 analysis is done with 18 queries instead of 20 and in total 98 queries were analyzed.



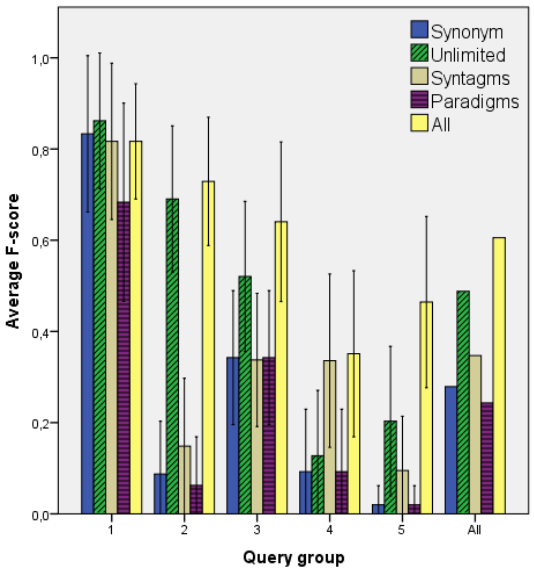
Graph 1. F-score Semantic Graph for High Quality.

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)=57.938$, $p<0.001$). Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a 0.01 level of significance (0.05/5 conditions). The results can be found in Table I.

Table I. F-score All Semantic Graph Wilcoxon for High Quality.

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-4.458, p<0.001*	Z=-3.128, p=0.02*	Z=-1.890, p=0.059	Z=-5.224, p<0.001*
UNL	Z=-4.458, p<0.001*		Z=-2.401, p=0.016*	Z=-4.859, p<0.001*	Z=-2.162, p=0.031*
SYNT	Z=-3.128, p=0.02*	Z=-2.401, P=0.016*		Z=-3.635, p<0.001*	Z=-3.985, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-4.859, p<0.001*	Z=-3.635, p<0.001*		Z=-5.635, p<0.001*
ALL	Z=-5.224, p<0.001*	Z=-2.162, p=0.031*	Z=-3.985, p<0.001*	Z=-5.635, p<0.001*	

The order of overall performance is thus SYNONYM = PARADIGM > SYNTAGM > UNLIMITED SEMIOSIS > ALL, all significant differences. For group 1 no significant differences between SYNONYM and the other methods are found. For group 2 significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-3.550, p<0.001), SYNTAGM (Z=-3.432, p=0.001) and PARADIGM (Z=-3.651, p<0.001) are found. For group 3 no significant differences between PARADIGM and the other methods are found. For group 4 significant differences between SYNTAGM and SYNONYM (Z=-2.670, p=0.008) and PARADIGM (Z=-2.670, p=0.008) are found. For group 5 significant differences between ALL and SYNONYM (Z=-3.053, p=0.002), SYNTAGM (Z=-2.833, p=0.005) and PARADIGM (Z=-3.053, p=0.002) are found.



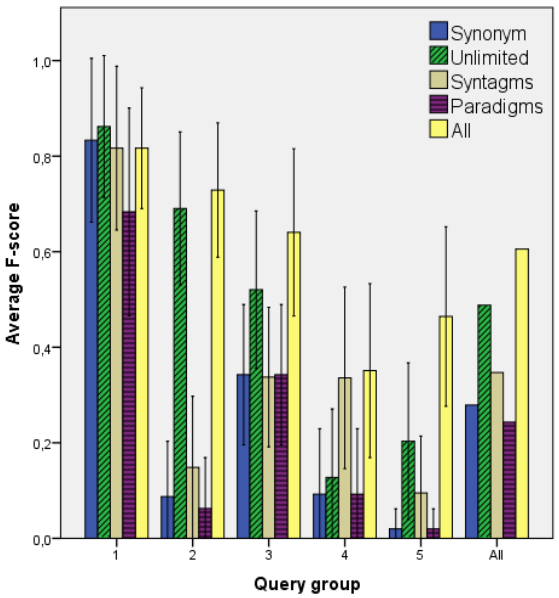
Graph 2. F-score Semantic Graph for Neutral.

2) Neutral system ($\beta = 1$): Graph 2 shows the F-score for the neutral system for each of the methods for each type of query group with the confidence interval of 95%.
A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)=98.571$, $p<0.001$). Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a 0.01 level of significance (0.05/5 conditions). The results can be found in Table II.

Table II. F-score All Semantic Graph Wilcoxon for Neutral.

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-5.050, p<0.001*	Z=-3.043, p=0.02*	Z=-1.890, p=0.059	Z=-6.049, p<0.001*
UNL	Z=-5.050, p<0.001*		Z=-3.276, p=0.001*	Z=-5.366, p<0.001*	Z=-3.535, p<0.001*
SYNT	Z=-3.043, p=0.02*	Z=-3.276, p=0.001*		Z=-3.576, p<0.001*	Z=-5.329, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-5.366, p<0.001*	Z=-3.576, p<0.001*		Z=-6.434, p<0.001*
ALL	Z=-6.049, p<0.001*	Z=-3.535, p<0.001*	Z=-5.329, p<0.001*	Z=-6.434, p<0.001*	

The order of overall performance is thus equal to the performance for the high quality system. The same significant differences are found for query group 1, 2, and 5. For group 3 significant differences between PARADIGM and UNLIMITED SEMIOSIS (Z=-2.805,p=0.005) and ALL (Z=-3.237, p=0.001) exist, as well as significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-2.805,p=0.005), PARADIGM (Z=-2.805,p=0.005) and SYNTAGM (Z=-2.926, p=0.003). For group 4 an additional significant difference between SYNTAGM and UNLIMITED SEMIOSIS (Z=-2.603, p=0.009) is found.



Graph 3. F-score Semantic Graph for High Effectiveness.

3) High effectiveness system ($\beta = 10$): Graph 3 shows the F-score for the high effectiveness system for each of the methods for each type of query group with the confidence interval of 95%.
A Friedman test showed a statistically significant difference among the methods, $\chi^2(4)=108.197$, $p<0.001$. Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a 0.01 level of significance (0.05/5 conditions). The results can be found in Table III.

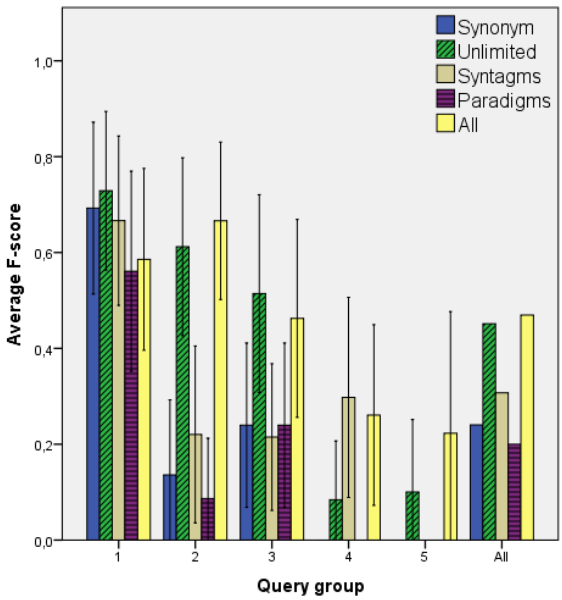
Table III . F-score All Semantic Graph Wilcoxon for High Effectiveness.

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-5.230, p<0.001*	Z=-3.241, p=0.001*	Z=-1.890, p=0.059	Z=-6.664, p<0.001*
UNL	Z=-5.230, p<0.001*		Z=-3.524, p<0.001*	Z=-5.521, p<0.001*	Z=-4.815, p<0.001*
SYNT	Z=-3.241, p=0.001*	Z=-3.524, p<0.001*		Z=-3.716, p<0.001*	Z=-6.083, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-5.521, p<0.001*	Z=-3.716, p<0.001*		Z=-6.896, p<0.001*
ALL	Z=-6.664, p<0.001*	Z=-4.815, p<0.001*	Z=-6.083, p<0.001*	Z=-6.896, p<0.001*	

The order of overall performance is thus equal to the performance for both the high quality and neutral system. The same significance values are found as for the neutral system, except for the significant difference between PARADIGM and ALL in group 3. This difference is no longer significant for the high effectiveness system.

B. *Image Retrieval*

1) High precision system ($\beta = 0.1$): Graph 4 shows the F-score for the high quality system for each of the methods



Graph 4. F-score Image Retrieval for High Quality.

for each type of query group with the confidence interval of 95%. For 14 queries of which one in group 1, three in group 4 and ten in group 5, no relevant annotation was available. In total 86 queries are analyzed.

A Friedman test showed a statistically significant difference among the methods, $\chi^2(4)=58.891$, $p<0.001$. Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a 0.01 level of significance (0.05/5 conditions). The results can be found in Table IV.

Table IV. F-score All Image Retrieval Wilcoxon for High Quality.

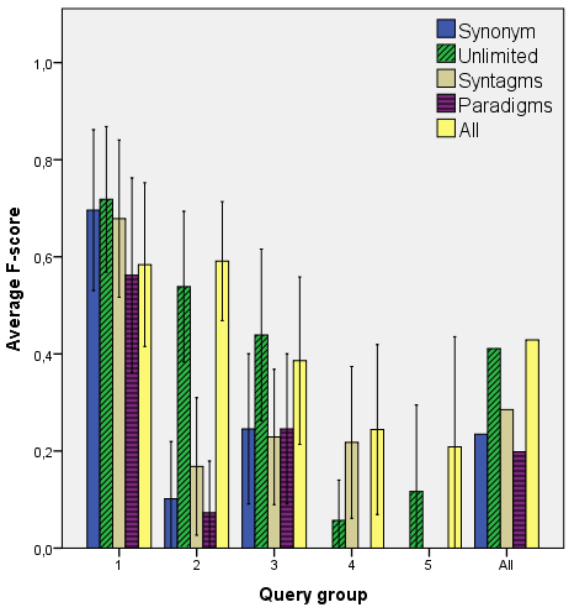
	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-4.684, p<0.001*	Z=-2.511, P=0.012*	Z=-1.826, p=0.068	Z=-4.108, p<0.001*
UNL	Z=-4.684, p<0.001*		Z=-3.060, p=0.002*	Z=-5.012, p<0.001*	Z=-0.686, p=0.493
SYNT	Z=-2.511, p=0.012*	Z=-3.060, p=0.002*		Z=-3.155, p=0.002*	Z=-3.250, p=0.001*
PARA	Z=-1.826, p=0.068	Z=-5.012, p<0.001*	Z=-3.155, p=0.002*		Z=-4.722, p<0.001*
ALL	Z=-4.108, p<0.001*	Z=-0.686, p=0.493	Z=-3.250, p=0.001*	Z=-4.722, p<0.001*	

The order of overall performance is thus SYNONYM = PARADIGM > SYNTAGM > UNLIMITED SEMIOSIS = ALL, all significant differences. For group 1 no significant differences between SYNONYM and the other methods are found. For group 2 significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-3.294, p=0.001), SYNTAGM (Z=-2.982, p=0.003) and PARADIGM (Z=-3.413, p=0.001) are found. For group 3 significant differences between PARADIGM and UNLIMITED SEMIOSIS (Z=-2.701, p=0.007) exist, as well as significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-2.701, p=0.007), PARADIGM (Z=-2.701, p=0.007) and SYNTAGM (Z=-2.845, p=0.004). For group 4 no significant differences between SYNTAGM and the other methods are found and for group 5 no significant differences were found.

2) Neutral system ($\beta = 1$): Graph 5 shows the F-score for the neutral system for each of the methods for each type of query group with the confidence interval of 95%.

Table V. F-score All Image Retrieval Wilcoxon for Neutral.

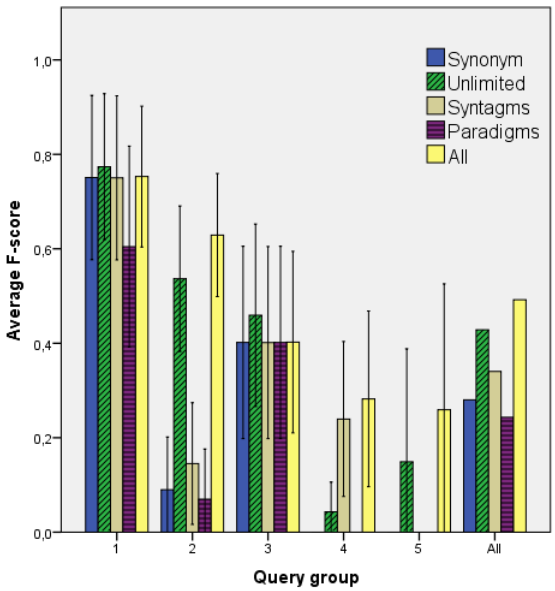
	SYNO	UNL	SYNT	PARA	ALL
SYNO		F=-4.723, p<0.001*	F=-2.354, p=0.019*	F=1.826, p=0.068	F=-4.019, p<0.001*
UNL	F=-4.723, p<0.001*		F=-3.350, p=0.001*	F=-5.045, p<0.001*	F=-0.800, p=0.424
SYNT	F=-2.354, p=0.019*	F=-3.350, p=0.001*		F=-3.052, p=0.002*	F=-3.554, p<0.001*
PARA	F=1.826, p=0.068	F=-5.045, p<0.001*	F=-3.052, p=0.002*		F=-4.704, p<0.001*
ALL	F=-4.019, p<0.001*	F=-0.800, p=0.424	F=-3.554, p<0.001*	F=-4.704, p<0.001*	



Graph 5. F-score Image Retrieval for Neutral.

A Friedman test showed a statistically significant difference among the methods, $\chi^2(4)=71.047$, $p<0.001$. Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at 0.01 level of significance (0.05/5 conditions). The results can be found in Table V. The same significant differences between conditions for ALL and the different query groups can be found as for the high quality system.

3) High effectiveness system ($\beta = 10$): Graph 6 shows the F-score for the high effectiveness system for each of the methods for each type of query group with the confidence interval of 95%.



Graph 6. F-score Image Retrieval for High Effectiveness.

A Friedman test showed a statistically significant difference among the methods, $\chi^2(4)=67.386$, $p<0.001$. Wilcoxon Signed-Ranks Test was used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a 0.01 level of significance (0.05/5 conditions). The results can be found in Table VI.

Table VI. F-score All Image Retrieval Wilcoxon for High Effectiveness.

	SYNO	UNL	SYNT	PARA	ALL
SYNO		F=-4.586, p<0.001*	F=-2.825, p=0.005*	F=-1.826, p=0.068	F=-4.775, p<0.001*
UNL	F=-4.586, p<0.001*		F=-2.654, p=0.008*	F=-4.930, p<0.001*	Z=-2.257, p=0.024*
SYNT	F=-2.825, p=0.005*	F=-2.654, p=0.008*		Z=-3.362, p=0.001*	Z=-4.184, p<0.001*
PARA	F=-1.826, p=0.068	F=-4.930, p<0.001*	Z=-3.362, p=0.001*		Z=-5.196, p<0.001*
ALL	F=-4.775, p<0.001*	Z=-2.257, p=0.024*	Z=-4.184, p<0.001*	Z=-5.196, p<0.001*	

The same significant differences between conditions for ALL and the different query groups can be found as for the high quality and neutral system, except that we have no longer significant differences in group 3.

VIII. DISCUSSION

In the discussion, we reflect on the experimental results regarding the use of semiotic structures to close the semantic gap in CBIR applications, both its semantic matching and its image retrieval parts. Additionally, we discuss the limitations of this research.

A. Semantic matching

Results on the semantic matching show that the use of the ALL method for query expansion, e.g., taking into account each and every type of relation that is available in the knowledge base, gives best overall performance independent of the type of application you want to use, i.e., high quality, neutral or high effectiveness. This effect is mainly rooted in the *RelatedTo* relation; this relation does not reflect any semiotic structure and was hence excluded from the other methods, however, it leads to alternative concepts that appear relevant to the original query concept. Examples of such relations produce expansions such as 'key fob' to 'key ring' and 'aircraft' to 'airplane', which can fuel a debate whether their relation with the query concept would not be better expressed as *synonym*. The method UNLIMITED SEMIOSIS gives the second best overall performance for all types of applications, both significantly lower than ALL and significantly higher than the other methods. As expected, this method turns the external knowledge base into a directed graph that expresses levels of aggregation, and therefore finds more abstract or more specific concepts compared to the baseline. The method SYNTAGMS has the third overall performance. This semiotic type is particularly good due to the fourth group of queries where a translation to another syntagmatic type is due for success, i.e., from *action* to *object* or from *attribute* to *object*. PARADIGMS have equal performance, or even slightly worse, than BASELINE. Whereas our hypothesis was that it would only exclude irrelevant concepts, it also excluded some concepts that were annotated as relevant, such as 'soccer_ball' for 'ball' and 'motorbike' for 'motorcycle'. Whether these relations should be excluded or included is based on the type of application that it is used for. Alternatives that are found by this semiotic type, and that are causing a decrease in performance, are all found as *synonym* as well. Hence, a strategy might be to include all relations that have the synonym relation independent of the presence of the PARADIGMS.

A closer look at the results for the different query groups as introduced in Section VI.B shows the following. For the first query group (*baseline*) no significant differences are found over the different expansion strategies. Surprisingly, no SYNONYM relation between 'key fob' and 'keyring', and between 'aircraft' and 'airplane' is available in ConceptNet. The first is present through a *RelatedTo* relation and the second through an *IsA* relation. Furthermore, no ConceptNet entry for 'camping bus' is present, so no expansions are found, dropping performance. Performance for PARADIGMS is lowest, because expansions for 'automobile' and 'beefburger', which are both considered relevant according to our ground truth, are paradigmatically excluded, dropping performance. SYNTAGMS is slightly lower than SYNONYM, because of an, as irrelevantly annotated, relation between 'shit' and 'cow', which is a debatable choice. UNLIMITED SEMIOSIS finds a relation between 'football' and 'skateboard', which slightly decreases performance. ALL has found several good alternatives, but also irrelevant ones, which is nicely visible in Graph 1 (relatively low F-score) and Graph 3 (relatively high F-score).

For the second query group (*unlimited semiosis*) the hypothesis was that UNLIMITED SEMIOSIS performs best. Significant differences between all other expansion methods, except ALL, are found. As the different graphs show, UNLIMITED SEMIOSIS has a better quality (Graph 1), whereas ALL has a higher effectiveness (Graph 2). The ALL method finds additional relevant concepts for 'animal' ('cow'), 'tool' ('screwdriver') and 'door' ('bus'), but irrelevant concepts for 'vehicle' ('tag') and 'leaf' ('pig'). The other methods find very little concepts and, therefore, performance is low.

The third query group (*paradigms*) is a group that expresses restrictions, such as spatial relation and color. No significant differences in performance are found for high precision applications, but for neutral and high effectiveness applications performance of UNLIMITED SEMIOSIS and ALL methods are significantly higher than PARADIGMS. This is due to relevant expansions for 'animal', 'flower', 'vehicle', 'hat', 'Mercedes' and 'Range Rover'. No results for 'air vehicle', 'water vehicle' and 'land vehicle' are found.

For the fourth query group (*syntagms*) the hypothesis was that the SYNTAGMS method performs best. As with the second query group and the UNLIMITED SEMIOSIS method, we see a high-quality for the SYNTAGMS method, but a high recall for the ALL method. The other methods have low performance, because they remain in the same syntagmatic part of the graph, whereas this query group requires a

transition to other syntagmatic alternatives. The main difference between SYNTAGMS and ALL is rooted in 'riding', 'stopping' and 'fast' in favor of ALL and 'landing' in favor of SYNTAGMS.

Finally, the fifth query group (*others*) are queries that have a very loose relation with the classifiers. Results show a significant difference between the ALL method and the other methods, as expected, but not compared to the UNLIMITED SEMIOSIS method. Many concepts in this group can only be found by ALL, but for 'headgear', 'tomato', 'farm' and 'wool', concepts are also found by the UNLIMITED SEMIOSIS method.

In the context of this case of the experiment, we can conclude that the type of query and the type of application prescribe the type of semiotic methods to consider. For applications that value effectiveness, the ALL method will be a good choice. Contrarily, for applications that require high quality, the UNLIMITED SEMIOSIS method would be a better choice, assuming that its queries do not require transitions between syntagmatic concepts (group 4), or are vaguely related to classifiers (group 5). Another good option for high quality applications would be to combine the SYNTAGMS and UNLIMITED SEMIOSIS methods. Finally, although in theory the PARADIGMS method should improve results for high quality applications, results indicate that it needs a more careful approach.

B. *Image retrieval*

Results on the image retrieval case show the impact of the semiotic structures on both parts of the semantic gap, and therefore the system as a whole. The general trend is that performance for this case is lower than for the semantic matching case, above. This originates from the fact that our classifiers do not perform very well. For instance, in query groups 4 (*syntagms*) and 5 (*other*) some expansion methods show no performance at all, which implies that despite the presence of relevant ground truth for them, none of the queries produce image results. The largest difference in overall performance between both cases is that the methods for UNLIMITED SEMIOSIS and ALL are no longer significantly different (Graphs 4 – 6). This is an indication that by adding irrelevant concepts (by the ALL method) more irrelevant images are produced, which might hurt more than adding less relevant concepts (by the UNLIMITED SEMIOSIS method) that produces less irrelevant images. This result even holds for high effectiveness applications.

A closer look at the results for the different query groups as introduced in Section VI.B shows the following. For the first group (*Synonyms*) not much difference is found over the various methods. Only the ALL methods drops a little more than the SYNTAGMS method, because the expansion by the ALL method from 'motorbike' to concepts 'horse' and 'helmet' really hurts performance as both are not synonyms while any image with either a horse, a helmet or a motorcycle will still be retrieved. As indicated above, the performance of the image retrieval case is lower than the semantic matching case. In this query group that is exemplified by the fact that although our classifiers for 'boat', 'motorcycle' and 'turd' are performing flawless, 'car', 'bus', 'traffic light' and 'turnscrew' perform less optimal ($F_{0.1}$ ~80%), whilst the classifiers for 'airplane', 'helmet' and 'football' can at best be graded acceptable ($F_{0.1}$ ~60%).

In the second query group (*unlimited semiosis*), the UNLIMITED SEMIOSIS method performs best. Significant differences between all other expansion methods, except ALL, are found. Differently from the results in the Semantic Matching, the UNLIMITED SEMIOSIS method is not better than ALL for high quality applications.

Results from the third query group (*paradigms*) interestingly show that the UNLIMITED SEMIOSIS method is slightly, but not significantly, better than its counterpart ALL, even for high recall applications. However, this is not only because of irrelevant expansions by the ALL method. In this group, many paradigmatic restrictions are specified by the queries, specifically about color, and colors cause a large decrease in performance in image retrieval. For example, the ALL method produces a semantic match between 'silver' and 'gray', indicating that gray cars are relevant. Unfortunately, in the image retrieval part silver cars are not detected as silver, but mainly as black. This is because many of the cars have black windows. Another example shows that green traffic lights are never detected, because the main color of the traffic light is black, irrespective of the light that is lit. In fact, this represents a typical example for unlimited semiosis where the semantic value of 'green' refers to an abstraction level that is far above the specific level that is indicated by the minimal part of the object that actually represents the green lit light. After all, we are not searching for a completely green traffic light. Besides the color classifiers, also other classifiers perform suboptimal, which has a negative effect on the results. Because, when a classifier is not able to detect the relevant concept in a relevant image, no difference between the methods can be registered.

Image retrieval results in the fourth group (*syntagms*) show similar results as in the semantic matching case: a slightly higher quality for the SYNTAGMS method and a higher effectiveness for the ALL method. These differences are, however, not significant any more. This is also the case for the fifth query group (*others*): no differences compared to semantic graph results, while the results are not significant anymore.

Overall, we can thus conclude that for high quality applications, the ALL method potentially hurts performance. Already for neutral applications, the UNLIMITED SEMIOSIS method, or a combined application of the SYNTAGMS and UNLIMITED SEMIOSIS methods might be a better choice than the ALL method. Additionally, this conclusion might prove stronger when taking into account the end user of the system whom might judge the results from the ALL method far worse than the results from the semiotic methods: in a retrieval system with many irrelevant results, as with application of the ALL method, it would be hard to find the relevant results amongst them, whilst the less, but more relevant results of the UNLIMITED SEMIOSIS method will be much easier to detect by the end user.

C. Limitations of experiment

One of the limitations of these experiments is that our dataset is really small. With only 51 classifiers, the probability that any of the words in ConceptNet matches our classifier labels is, therefore, much lower. Then, one single true positive has a major impact on score whilst the many false positives that happen to have no match do not add to the score balance. This might be the reason that the ALL method is performing better than we expected.

A second limitation is performance of the classifiers. As explained in part B. of this section, our color classifiers as well as some object classifiers are suboptimal. In order to profit from improvements in the semantic reasoning part of the system, good classifiers are needed. This argument also holds in reverse: on optimizing classifiers, overall little will be gained unless the improvements in this part of the semantic gap is matched with an equal improvement in the semantic matching part of the semantic gap.

An algorithm performs only as good as the quality of the data it is provided with. Especially when the focus is on generic semantic knowledge, a third limitation is the knowledge base of choice. ConceptNet has a lot of different types of relations and, therefore, connections between concepts exists that apply different relations than expected, i.e., impacting accuracy, or no relations are available at all where one would expect their occurrence, impacting completeness. Although we experienced major improvements of version 5.3 over 5.2, e.g., corrections from erroneous relationships, several flaws in our experiment find their root in debatable concept relations from ConceptNet, or absent concepts. Another lesson learned from ConceptNet is the use of underscored words. Underscored words represent complex concepts that are represented by composition of two or more words by applying underscores, e.g., 'woman_wardrobe' or 'red_traffic_light'. Humans easily recognize their (syntagmatic) structure, but putting such understanding into (semiotic) rules is another matter completely. Therefore, we decided to abandon their use altogether, in order to stay away from potentially incorrect expansion results from factually correct data such as *CapableOf(camper, shoe_away_bear)* and *PartOf(dress, woman_wardrobe)*.

Finally, we have designed the experiment to score against two ground truths, one for the semantic matching and one for the image retrieval. They have the 100 queries in common, and since we have only 20 queries for each query group (Section VI.B) they also share their susceptibility to annotation-induced performance variations. We acknowledge this weakness in our experiment, especially since each annotation is performed by one individual each.

IX. CONCLUSION AND FUTURE WORK

In conclusion, applying semiotic relations in query expansion over an external, generic knowledge base, contributes to a higher quality semantic match between query concepts and classifier labels, and also significantly improves image retrieval performance compared to a baseline with only synonym expansions. The type of query and the type of application prescribe the type of semiotic methods that should be considered for semantic matching. The indiscriminate use of all available relations that are present in the external knowledge base potentially hurts performance of the image retrieval part. The same approach for the semantic matching surprisingly outperformed the dedicated semiotic methods, although we have strong reasons to believe this effect is rooted in coincidental flaws in the knowledge base of choice. The experiment results also confirmed that the semantic gap that is experienced within CBIR consists of two cascading parts, and that little is gained overall when improvements address one part only. Finally, although multiple relations from the external knowledge base have been mapped onto one single semiotic method that at best approximates the semantics of the underlying relations, it is above doubt that semiotic coherence emerges in the otherwise non-semiotic semantic network that the external knowledge base represents. We have shown that this semiotic coherence can be employed to improve the semantic capability of a software system.

In future research, it is advisable to explore the effectiveness of these semiotic structures on other knowledge bases, containing either generic or domain-specific knowledge, in order to further evaluate the true genericity of this semiotic approach. Specifically related to ConceptNet it may be worthwhile to investigate appropriate (semiotic) ways to handle complex concepts (underscored words) in order to disclose their knowledge and improve query expansion.

Inclusion of more classifiers, including better color classifiers, and more classifier types, such as action classifiers and object relation classifiers, will improve the significance of the outcome of the experiments as well as the applicability of the expansion methods.

Furthermore, it would be interesting to conduct research into the influence of other semiotic structures, such as the semiotic square about contradictions, expressing relations that are also available in external databases, e.g., negated concepts and antonyms.

Additionally, it would be beneficial to measure image retrieval performance using relevance feedback from an end user on the found classifier labels by ConceptNet. For instance, our use of paradigms is completely unaware of the intentions of the end user and therefore might wrongly exclude a specific set of paradigmatic concepts. This can be easily corrected by adding context of use through relevance feedback.

ACKNOWLEDGMENT

This research was performed in the context of the GOOSE project, which is jointly funded by the enabling technology program Adaptive Multi Sensor Networks (AMSN) and the MIST research program of the Dutch Ministry of Defense. Furthermore, we acknowledge the ERP Making Sense of Big Data for their financial support. The authors want to express their gratitude to Dignée Brandt for her annotation of the ground truth, and to Charelle Bottenheft for her support to the statistical analysis.

REFERENCES

[1] M. H. T. de Boer, L. Daniele, P. Brandt, and M. Sappelli, "Applying Semantic Reasoning in Image Retrieval," in *ALLDATA 2015, The 1st Int. Conf. on Big Data, Small Data, Linked Data and Open Data*, 2015, pp. 69–74.

[2] K. Schutte et al., "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation," in *CBMI*, 2015, pp. 1–4.

[3] K. Schutte et al., "GOOSE: Semantic search on internet connected sensors," in *SPIE Defense, Security, and Sensing*, 2013, p. 875806.

[4] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[5] P. G. B. Enser, C. J. Sandom, J. S. Hare, and P. H. Lewis, "Facing the reality of semantic image retrieval," *J. Doc.*, vol. 63, no. 4, pp. 465–481, 2007.

[6] H. Bouma, P. T. Eendenbak, K. Schutte, G. Azzopardi, and G. J. Burghouts, "Incremental concept learning with few training examples and hierarchical classification," *Proc. SPIE*, vol. 9652, 2015.

[7] J. Schavemaker, M. Spitters, G. Koot, and M. de Boer, "Fast re-ranking of visual search results by example selection," in *CAIP 2015, Part I, LNCS 9256*, 2015, pp. 387–398.

[8] R. Speer and C. Havasi, "Representing General Relational Knowledge in {ConceptNet} 5.," in *LREC*, 2012, pp. 3679–3686.

[9] K. Schutte et al., "TOSO dataset." TNO, The Hague, The Netherlands, 2015.

[10] M. H. T. de Boer et al., "Ground Truth on 100 User Queries about TOSO Dataset," 2015. [Online]. Available: DOI:10.13140/RG.2.1.3688.9049.

[11] J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[12] P. Perona, "Vision of a Visipedia," *Proc. IEEE*, vol. 98, no. 8, pp. 1526–1534, 2010.

[13] L. Bai, S. Lao, G. J. F. Jones, and A. F. Smeaton, "Video semantic content analysis based on ontology," in *IMVIP 2007*, 2007, pp. 117–124.

[14] A. D. Bagdanov, M. Bertini, A. Del Bimbo, G. Serra, and C. Torniai, "Semantic annotation and retrieval of video events using multimedia ontologies," in *ICSC 2007*, 2007, pp. 713–720.

[15] A. Oltramari and C. Lebiere, "Using ontologies in a cognitive-grounded system: automatic action recognition in video surveillance," in *Proc. Int. Conf. STIDS*, 2012.

[16] G. Erozal, N. K. Cicekli, and I. Cicekli, "Natural language querying for video databases," *Inf. Sci. (Ny).*, no. 178, pp. 2534–2552, Apr. 2008.

[17] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from web data," in *CICCV 2013*, 2013, pp. 1409–1416.

[18] M. de Boer, K. Schutte, and W. Kraaij, "Knowledge based query expansion in complex multimedia event detection," *Multimed. Tools Appl.*, pp. 1–19, 2015.

[19] D. D. K. Sleator and D. Temperley, "Parsing English with a link grammar," *arXiv Prepr. C.*, 1995.

[20] M. Nazrul Islam, "A systematic literature review of semiotics perception in user interfaces," *J. Syst. Inf. Technol.*, vol. 15, no. 1, pp. 45–77, Mar. 2013.

[21] D. Lamas and H.-L. Pender, "Reflection on the role of semiotic engineering in co-design of interaction," *IEEE Lat. Am. Trans.*, vol. 12, no. 1, pp. 48–53, Jan. 2014.

[22] F. Nicastro et al., "A semiotic-informed approach to interface guidelines for mobile applications: A case study on phenology data acquisition," in *ICEIS 2015*, 2015, vol. 3, pp. 34–43.

[23] B. Shishkov, J. L. G. Dietz, and K. Liu, "Bridging the language-action perspective and organizational semiotics in SDBC," in *ICEIS 2006*, 2006, pp. 52–60.

[24] C. M. de A. Barbosa, R. O. Prates, and C. S. de Souza, "MArq-G*," in *CLIHC '05*, 2005, vol. 124, pp. 128–138.

[25] A. L. O. Paraense, R. R. Gudwin, and R. de Almeida Goncalves, "Brainmerge: a Semiotic-Oriented Software Development Process for Intelligence Augmentation Systems," in *2007 Int. Conf. on Integration of Knowledge Intensive Multi-Agent Systems*, 2007, pp. 261–266.

[26] B. Shishkov, M. van Sinderen, and B. Tekinerdogan, "Model-Driven Specification of Software Services," in *ICEBE'07*, 2007, pp. 13–21.

[27] C. Landauer, "Layers of Languages for Self-Modeling Systems," in *IEEE SASO 2011*, 2011, pp. 214–215.

[28] M. Kwiatkowska, K. Michalik, and K. Kielan, *Soft Computing in Humanities and Social Sciences*, vol. 273. Springer Berlin Heidelberg, 2012.

[29] S. Yi'an, "A survey of peirce semiotics ontology for artificial intelligence and a nested graphic model for knowledge representation," in *CEUR Workshop Proc.*, 2013, vol. 1126, pp. 9–18.

[30] S. Chai-Arayalert and K. Nakata, "Towards a semiotic approach to practice-oriented knowledge transfer," in *Proc. KMIS 2012*, 2012, pp. 119–124.

[31] S. Liu, W. Li, and K. Liu, "Pragmatic Oriented Data Interoperability for Smart Healthcare Information Systems," in *2014 14th IEEE/ACM Int.Symp. on Cluster, Cloud and Grid Computing*, 2014, pp. 811–818.

[32] J. Yoon, "Improving recall of browsing sets in image retrieval from a semiotics perspective," Ph.D. dissertation, University of North Texas, 2006.

[33] E. Hartley, "Bound together: Signs and features in multimedia content representation," in *Proc. COSIGN 2004*, 2004.

[34] D. Chandler, *Semiotics: the basics*, 2nd ed. Routledge, 2007.

[35] C. S. Peirce, *Collected Papers of Charles Sanders Pierce: Elements of logic*, vol. 2. Cambridge, MA: Harvard University Press, 1932.

[36] U. Eco, *A theory of semiotics*. Bloomington, IN: Indiana University Press / London: Macmillam, 1976.

[37] D. Silverman and B. Torode, *The Material Word: Some Theories of Language and its Limits*, In: London: Routledge & Kegan Paul, 1980.

[38] M.-C. De Marneffe, B. MacCartney, C. D. Manning, and others, "Generating typed dependency parses from phrase structure parses," in *Proc. LREC*, 2006, vol. 6, no. 2006, pp. 449–454.

[39] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, 2009.