VigNet: Grounding Language in Graphics using Frame Semantics

Bob Coyne and **Daniel Bauer** and **Owen Rambow**
Columbia University
New York, NY 10027, USA
{coyne, bauer, rambow}@cs.columbia.edu

Abstract

This paper introduces *Vignette Semantics*, a lexical semantic theory based on Frame Semantics that represents conceptual and graphical relations. We also describe a lexical resource that implements this theory, VigNet, and its application in text-to-scene generation.

1 Introduction

Our goal is to build a comprehensive text-to-graphics system. When considering sentences such as *John is washing an apple* and *John is washing the floor*, we discover that rather different graphical knowledge is needed to generate static scenes representing the meaning of these two sentences (see Figure 1): the human actor is assuming different poses, he is interacting differently with the thing being washed, and the water, present in both scenes, is supplied differently. If we consider the types of knowledge needed for scene generation, we find that we cannot simply associate a single set of knowledge with the English verb *wash*. The question arises: how can we organize this knowledge and associate it with lexical items, so that the resulting lexical knowledge base both is usable in a wide-coverage text-to-graphics system, and can be populated with the required knowledge using limited resources?

In this paper, we present a new knowledge base that we use for text-to-graphics generation. We distinguish three types of knowledge needed for our task. The first is **conceptual knowledge**, which is knowledge about concepts, often evoked by words. For example, if I am told John bought an apple, then I know that that event necessarily also involved the seller and money. Second, we need **world knowledge**. For example, apples grow on trees in certain geographic locations at certain times of the year. Third, we need **grounding knowledge**, which tells us how concepts are related to sensory experiences. In our application, we model grounding knowledge with a database of 3-dimensional graphical models. We will refer to this type of grounding knowledge as **graphical knowledge**. An example of grounding knowledge is knowing that several specific graphical models represent apple trees.

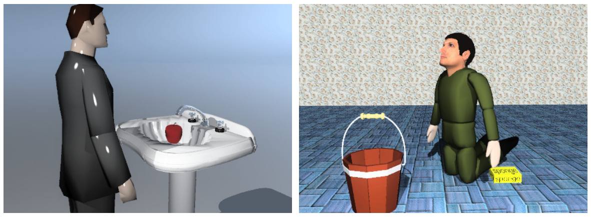

Figure 1: Mocked-up scenes using the WASH-SMALL-FRUIT vignette ("John washes the apple") and WASH-FLOOR-W-SPONGE vignette ("John washes the floor").

Conceptual knowledge is already the object of extensive work in frame semantics; FrameNet (Ruppenhofer et al., 2010) is an extensive (but not complete) relational semantic encoding of lexical meaning in a frame-semantic conceptual framework. We use this prior work, both the theory and the resource, in our work. The encoding of world knowledge has been the topic of much work in Artificial Intelligence. Our specific contribution in this paper is the integration of the representation for world knowledge and graphical knowledge into a frame-semantic approach. In order to integrate these knowledge types, we extend FrameNet in three manners.

1. Frames describe complex relations between their frame elements, but these relations, i.e.

Proceedings of the ACL 2011 Workshop on Relational Models of Semantics (RELMS 2011), pages 28–36,
Portland, Oregon, USA, June 23, 2011. ©2011 Association for Computational Linguistics

the internal structure of a frame, is not explicitly formulated in frame semantics. FrameNet frames do not have any intensional meaning besides the informal English definition of the frames (and what is expressed by so-called “frame-to-frame relations”). From the point of view of graphics generation, internal structure is necessary. While for many applications a semantic representation can remain vague, a scene must contain concrete objects and spatial relations between them.

2. Some frames are not semantically specific enough. For example, there is a frame SELF_MOTION, which includes both *walk* and *swim*; these verbs clearly need different graphical realizations, but they are also different from a general semantic point of view. While this situation could be remedied by extending the inventory of frames by adding WALK and SWIM frames, which would inherit from SELF_MOTION, the situation is more complex. Consider *wash an apple* and *wash the floor*, discussed above. While the core meaning of *wash* is the same in both phrases, the graphical realization is again very different. However, we cannot simply create two new frames, since at some level (though not the graphical level) the meaning is indeed compositional. We thus need a new mechanism.

3. FrameNet is a lexical resource that illustrates how language can be used to refer to frames, which are abstract definitions of concepts, and their frame elements. It is not intended to be a formalism for deep semantic interpretation. The FrameNet annotations show the frame elements of frames (e.g. the goal frame element of the SELF_MOTION frame) being filled with text passages (e.g. *into the garden*) rather than with concrete semantic objects (e.g. an ‘instance’ of a LOCALE_BY_USE frame evoked by *garden*). Because such objects are needed in order to fully represent the meaning of a sentence and to assert world knowledge, we introduce **semantic nodes** which are discourse referents of lexical items (whereas frames describe their meanings).

In this paper, we present VigNet, a resource which extends FrameNet to incorporate world and graphical knowledge. We achieve this goal by addressing the three issues above. We first extend frames by adding more information to them (specifically, about decomposition relevant to graphical grounding and more precise selectional restrictions). We call a frame with graphical information a **vignette**. We then extend the structure defined by FrameNet by adding new frames and vignettes, for example for *wash an apple*. The result we call VigNet. Finally, we extend VigNet with a system of nodes which instantiate frames; these nodes we call **semantic nodes**. They get their meaning only from the frames they instantiate. All three extensions are conservative extensions of frames and FrameNet. The semantic theory that VigNet instantiates we call **Vignette Semantics** and we believe it to be a conservative extension (and thus in the spirit of) frame semantics.

This paper is structured as follows. In Section 2, we review frame semantics and FrameNet. Section 3 presents a more detailed description of VigNet, and we provide examples in Section 4. Since VigNet is intended to be used in a large-coverage system, the population of VigNet with knowledge is a crucial issue which we address in Section 5. We discuss related work in Section 6 and conclude in Section 7.

2 Frame Semantics and FrameNet

Frame Semantics (FS; Fillmore (1982)) is based on the idea that the meaning of a word can only be fully understood in context of the entire conceptual structure surrounding it, called the word’s frame. When the meaning of a word is evoked in a hearer’s mind all related concepts are activated simultaneously and we can rely on this structure to transfer information in a conversation. Frames can describe states-of-affairs, events or complex objects. Each frame contains a set of specific frame elements (FEs), which are labeled semantic argument slots describing participants in the frame. For instance, the word *buy* evokes the frame for a commercial transaction scenario, which includes a buyer and a seller that exchange money for goods. A speaker is aware of what typical buyers, sellers, and goods are. He may also have a mental prototype of the visual scenario itself (e.g. standing at a counter in a store). In FS the role of syntactic theory and the lexicon is to explain how the syntactic dependents of a word that realizes a frame (i.e. arguments and adjuncts) are mapped to frame elements via valence patterns.

FrameNet (FN; Baker et al. (1998), Ruppenhofer et al. (2010)) is a lexical resource based on FS. Frames in FN (around 1000) [1] are defined in terms of their frame elements, relations to other frames and semantic types of FEs. Beyond this, the meaning of the frame (how the FEs are related to each other) is only described in natural language. FN contains about 11,800 lexical units, which are pairings of words and frames. These come with annotated example sentences (about 150,000) to illustrate their valence patterns. FN contains a network of directed frame-to-frame relations. In the INHERITANCE relation a child-frame inherits all semantic properties from the superframe. The frame relations SUBFRAME and PRECEDES refer to sub-events and events following in temporal order respectively. The parent frame's FEs are mapped to the child's FEs. For instance CAUSE_TO_WAKE inherits from TRANSITIVE_ACTION and its sleeper FE maps to agent. Other relations include PERSPECTIVE_ON, CAUSATIVE_OF, and INCHOATIVE_OF. Frame relations captures important semantic facts about frames. For instance the hierarchical organization of INHERITANCE allows to view an event on varying levels of specificity. Finally, FN contains a small ontology of semantic types for frame elements, which can be interpreted as selectional restrictions (e.g. an agent frame element must be filled by a **sentient** being).

3 Vignette Semantics

In Section 1, we motivated VigNet by the need for a resource that allows us to relate language to a grounded semantics, where for us the graphical representation is a stand-in for grounding. We described three reasons for extending FrameNet to VigNet: we need more meaning in a frame, we need more frames and more types of frames, and we need to instantiate frames in a clean manner. We discuss these refinements in more detail in this section.

- **Vignettes** are frames that are decomposed into graphical primitives and can be visualized. Like other fames they are motivated by frame semantics; they correspond to a conceptual structure evoked by the lexical units which are associated with it.

- VigNet includes individual frames for each (content) lexical item. This provides **finer-grained semantics** than given with FrameNet frames themselves. These lexically-coupled frames leverage the existing structure of their parent frames. For example, the SELF_MOTION frame contains lexical items for *run* and *swim* which have very different meaning even though they share the same frame and FEs (such as SOURCE, GOAL, and PATH). We therefore define frames for RUN and SWIM which inherit from SELF_MOTION. We assume also that frames and lexical items that are missing from FrameNet are defined and linked to the rest of FrameNet as needed.

- Even more specific frames are created to represent **composed vignettes**. These are vignettes that ground meaning in different ways than the primitive vignette that they specialize. The only motivation for their existence is the graphical grounding. For example, we cannot determine how to represent washing an apple from the knowledge of how to represent generic washing and an apple. So we define a new vignette specifically for *washing a small fruit*. From the point of view of lexical semantics, it uses two lexical items (wash and apple) and their interpretation, but for us, since we are interested in grounding, it is a single vignette. Note that it is not necessary to create specific vignettes for every concrete verb/argument combination. Because vignettes are visually inspired relatively few general vignettes (e.g. *manipulate an object on a fixture*) suffices to visualize many possible scenarios.

- A new type of frame-to-frame relation, which we call SUBFRAME-PARALLEL is used to decompose vignettes into a set of more primitive semantic relations between their arguments. Unlike FrameNet's SUBFRAME relation which

[1] Numbers refer to FrameNet 1.5

represents temporally sequential subframes, in SUBFRAME-PARALLEL, the subframes are all active at the same time, provide a conceptual and spatial decomposition of the frame, and can serve as spatial constraints on the frame elements. A frame is called a vignette if it can be decomposed into graphical primitives using SUBFRAME-PARALLEL relations. For instance in the vignette WASH-SMALL-OBJ for *washing a small object in a sink*, the washer has to be in front of the sink. We assert a SUBFRAME-PARALLEL relation between WASH-SMALL-OBJ and FRONTOF, mapping the washer FE to the figure FE and sink to ground.

- FrameNet has a very limited number of semantic types that are used to restrict the values of FEs. Vignette semantics uses **selectional restrictions** to differentiate between vignettes that have the same parent. For example, the vignette invoked for washing a small object in a sink would restrict the semantic type of the theme (the entity being washed) to anything small, or, more generally, to any object that is washed in this way (apples, hard-boiled eggs, etc). The vignette used for washing a vehicle in a driveway with a hose would restrict its theme to some set of large objects or vehicle types. Selectional restrictions are asserted using the same mechanism as decompositions.

- As mentioned in Section 1, in FrameNet annotations frame elements (FEs) are filled with text spans. Therefore, while frame semantics in general is a deep semantic theory, FrameNet annotations only represent shallow semantics and it is not immediately obvious how FrameNet can be used to build a full semantic representations of a sentence. In Vignette semantics, when a frame is evoked by a lexical item, it is *instantiated* as a semantic node. Its FEs are then bound not to subphrases, but to semantic nodes which are the instantiations of the frames evoked by those subphrases.

Section 3.1 investigates semantic nodes in more detail. Section 3.2 illustrates different types of vignettes (objects, actions, locations) and how they are defined using the SUBFRAME_PARALLEL relation. In Section 3.3 we discuss selectional restrictions.

3.1 Semantic Nodes and Relational Knowledge

The intuition behind semantic nodes is that they represent objects, events or situations. They can also represent plurals or generics. For instance we could have semantic node **city**, denoting the class of cities and a semantic node **paris**, that denotes the city Paris. Note that there is also a frame CITY and a frame PARIS that contain the conceptual structure associated with the words *city* and *Paris*. Frames represent the linguistic and the conceptual aspect of knowledge; the intensional *meaning* of a word. They provide knowledge to answer questions such as “What is an apple?” or “How do you wash an apple?”. In contrast, semantic nodes are extensional, i.e. *denotations*. They represent the knowledge to answer questions such as “In what season are apples harvested?” or “How did Percy wash that apple just now?”.

As mentioned above semantic nodes allow us to build full meaning representations of entire sentences in discourse. Therefore, while frame definitions are fixed, semantic nodes can be added dynamically during discourse understanding or generation to model the instances of frames that language is evoking. We call such nodes **temporary semantic nodes**. They they are closely related to the discourse referents of Discourse Representation Theory (Kamp, 1981) and related concepts in other theories. In contrast, **persistent semantic nodes** are used to store world knowledge which is distinct from the conceptual knowledge encoded within frames and their relations; for example, the frame for *moon* will not encode the fact that the moon’s circumference is 6,790 miles, but we may record that using a knowledge based of external assertions semantic nodes are given their meaning by corresponding frames (CIRCUMFERENCE, MILE, etc.). A temporary semantic node can become persistent by being retained in the knowledge base.

3.2 Vignette Types and their Decomposition

A vignette is a frame in the FrameNet sense that is decomposed to a set of more primitive frames using the SUBFRAME-PARALLEL frame-to-frame relation. The frame elements (FEs) of a vignette are defined as in FrameNet, except that our grounding in the graphical representation gives us a new, strong criterion to choose what the FEs are: they are the objects necessarily involved in the visual scene associated with that vignette. The subframes represent the spatial and other relations between the FEs. The resulting semantic relations specify how the scene elements are spatially arranged. This mechanism covers several different cases.

For **actions**, we conceptually freeze the action in time, much as in a comic book panel, and represent it in a vignette with a set of objects, spatial relations between those objects, and poses characteristic for the humans (and other pliable beings) involved in that action. Action vignettes will typically be specialized to composed vignettes, so that the applicability of different vignettes with the same parent frame will depend on the values of the FEs of the parent. In the process of creating composed vignettes, FEs are often added because additional objects are required to play auxiliary roles. As a result, the FEs of an action vignette are the union of the semantic roles of the important participants and props involved in that enactment of the action with the FEs of the parent frame. For instance the following vignette describes one concrete way of *washing a small fruit*. Note that we have included a new FE sink which is not motivated in the frame WASH.[2] Note also that this vignette also contains a selectional restriction on its theme, which we will discuss in the next subsection and which is not shown here.

WASH-SMALL-FRUIT(washer, theme, sink)

FRONTOF(figure:washer, figure:sink)
FACING(figure:washer, figure:sink)
GRASP(grasper:washer, theme:theme)
REACH(reacher:washer, target:sink)

In this notation the head row contains the vignette name and its FEs in parentheses. For readability we will often omit FEs that are part of the vignette but not restricted or used in any mentioned relation. The lower box contains the vignette decomposition and implicitly specifies SUBFRAME-PARALLEL frame-to-frame relations. In the decomposition of a vignette V we use the notation F(a:b, $\cdots$) to indicate that the FE a of frame F is mapped to the FE b of V. When V is instantiated the semantic node binding to a must also be able to bind to b in F.

Locations are represented by vignettes which express constraints between a set of objects characteristic for the given location. The FEs of location vignettes include these constituent objects. For example, one type of living room (of many possible ones) might contain a couch, a coffee table, and a fireplace in a certain arrangement.

LIVING-ROOM_42(left_wall, far_wall, couch, coffee_table, fireplace)

TOUCHING(figure:couch, ground:left_wall)
FACING(figure:couch, ground:right_wall)
FRONTOF(figure:coffee_table, ground: sofa)
EMBEDDED(figure:fire-place, ground:far_wall)

Even ordinary **physical objects** will have certain characteristic parts with size, shape, and spatial relations that can be expressed by vignettes. For example, an object type such as a kind of *stop sign* can be defined as a two-foot-wide, red, hexagonal metal sheet displaying the word "STOP" positioned on the top of a 6 foot high post.

STOP-SIGN(sign-part, post-part, texture)

MATERIAL(theme:sign-part, material:METAL)
MATERIAL(theme:post-part, material:METAL)
DIAMETER(theme:sign-part, diameter:**2 feet**)
HEIGHT(theme:post-part, height:**6 feet**)
ONTOP(figure:sign-part, ground:post-part)
TEXTURE(theme:sign-part, texture:**"*STOP*"**)

In addition, many real-world objects do not correspond to lexical items but are elaborations on them or combinations. These **sublexical entities** can be represented by vignettes as well. For example, one such 3D object in our text-to-scene system is a goat head mounted on a piece of wood. This object is represented by a vignette with two FEs (ghead, gwood) representing the goat's head and the wood. The vignette decomposes into ON(ghead, gwood).

While there can be many vignettes for a single lexical item, representing the many ways a location, action, or object can be constituted, vignettes need not be specialized for every particular situation and can be more or less general. In one exteme creating vignettes for every verb/argument combination would clearly lead to a combinatorial explosion and is not feasible. In the other extreme we can define rather general vignettes. For example, a vignette

[2] FrameNet does not currently contain a WASH frame, but if it did, it would not contain an FE sink.

USE-TOOL for using a tool on a theme can be represented by the user GRASPING the tool and REACHING towards the theme. These vignettes can be used in decompositions of more concrete vignettes (e.g. HAMMER-NAIL-INTO-WALL). They can also be used directly if no other more concrete vignette can be applied (because it does not exist or its selectional restrictions cannot be satisfied). In this way by defining a small set of such vignettes we can visualize approximate scenes for a large number of descriptions.

3.3 Selectional Restrictions on Frame Elements

To define a frame we need to specify selectional restrictions on the semantic type of its FEs. Instead of relying on a fixed inventory of semantic types, we assert conceptual knowledge and external assertions over persistent semantic types. This allows us to use VigNet's large set of frames to represent such knowledge. For example, an *apple* can be defined as a small round fruit.

APPLE(self)

SHAPEOF(figure:self, shape:**spherical**)
SIZEOF(figure:self, size:**small**)

APPLE is simply a frame that contains a self FE, which allows us to make assertions about the concept (i.e. about any semantic node bound to the self FE). Frame elements of this type are not unusual in FrameNet, where they are mainly used for frames containing common nouns (for instance the Substance FE contains a substance FE). In VigNet we implicitly use self in all frames, including frames describing situations and events.

We use the same mechanism to define specialized compound vignettes such as WASH_SMALL_FRUIT. We extend WASH in the following way to restrict it to small fruits (we abreviate F(self:a) as a=F for readability).

WASH-SMALL-FRUIT(washer, theme, sink)

% selectional restrictions
sink=SINK, washer=PERSON,
theme=x, x=FRUIT,
SIZEOF(figure:x,size:**small**)

% decomposition
FRONTOF(figure:washer, figure:sink)
FACING(figure:washer, figure:sink)
GRASP(grasper:washer, theme:theme)
REACH(reacher:washer, target:sink)

4 Examples

In this section we give further examples of visual action vignettes for the verb *wash*. The selectional restrictions and graphical decomposition of these vignettes vary depending on the type of object being washed. The first example shows a vignette for *washing a vehicle*.

WASH-VEHICLE(washer, theme, instr, location)

washer=PERSON, theme=VEHICLE,
instr=HOSE, location=DRIVEWAY

ONSURFACE(figure:theme, ground:location)
FRONTOF(figure:washer, ground:theme)
FACING(figure:washer, ground:theme)
GRASP(grasper:washer, theme:instrument)
AIM(aimer:washer, theme:instr, target:theme)

The following two vignettes represent a case where the object being washed alone does not determine which vignette to apply. If the instrument is unspecified one or the other could be used. We illustrate one option in figure 1 (right).

WASH-FLOOR-W-SPONGE(washer,theme,instr)

washer=PERSON, theme=FLOOR,
instr=SPONGE

KNEELING(agent:washer), GRASP(grasper:washer, theme:instr), REACH(reacher:washer, target:theme)

WASH-FLOOR-W-MOP(washer, theme, instr)
washer=PERSON, theme=FLOOR, instr=MOP
GRASP(grasper:washer, theme:instr), REACHWITH(reacher:washer, target:theme, instr:instr)

It is easy to come up with other concrete vignettes for *wash* (washing windows, babies, hands, dishes...). As mentioned in section 3.2 more general vignettes can be defined for very broad object classes. In choosing vignettes, the most specific will be used (looking at type matching hierarchies), so general vignettes will only be chosen when more specific ones are unavailable. The following generic vignette describes *washing any large object*.

WASH-LARGE-OBJECT(washer, theme instrument)

washer=PERSON, theme=OBJECT,
instrument=SPONGE,
SIZEOF(figure:theme, size:**large**)

FACING(figure:washer, ground:theme)
GRASP(grasper:washer, theme:instrument)
REACH(reacher:washer, target:theme)

In our final example, a vignette for *picking fruit* uses the following assertion of world knowledge about particular types of fruit and the trees they come from:
SOURCE-OF(theme:x, source:y), APPLE(self:x), APPLETREE(self:y)
In matching the vignette to the verb frame and its arguments, the source frame element is bound to the type of tree for the given theme (fruit).

PICK-FRUIT(picker, theme, source)

picker=PERSON, theme=FRUIT, source=TREE,
SOURCEOF(theme:theme, source:source)

UNDERCANOPY(figure:picker, canopy:source)
GRASP(grasper:picker, theme:theme)
REACH(reacher:picker, target:source.branch)

5 VigNet

We are developing VigNet as a general purpose resource, but with the specific goal of using it in text-to-scene generation. In this section we first describe various methods to populate VigNet. We then sketch how we create graphical representations from VigNet meaning representations.

5.1 Populating VigNet

VigNet is being populated using several approaches:

- Amazon Mechanical Turk is being used to acquire scene elements for location and action vignettes as well as the spatial relations among those elements. For locations, Turkers are shown representative pictures of different locations as well as variants of similar locations, thereby providing distinct vignettes for each location. We also use Mechanical Turk to acquire general purpose relational information for objects and actions such as default locations, materials, contents, and parts.

- We extract relations such as typical locations for actions from corpora based on co-occurance patterns of location and action terms. This is based on ideas described in (Sproat, 2001). We also rely on corpora to induce new lexical units and selectional preferences.

- A large set of semantic nodes and frames for nouns has been imported from the noun lexicon of the WordsEye text-to-scene system (Coyne and Sproat, 2001). This lexicon currently contains 15,000 lexical items and is tied to a library of 2,200 3D objects and 10,000 images Semantic relations between these nodes include parthood, containment, size, style (e.g. antique or modern), overall shape, material, as well as spatial tags denoting important spatial regions on the object. We also import graphically-oriented vignettes from WordsEye. These are used to capture the meaning of sub-lexical 3D objects such as the mounted goat head described earlier.

- Finally, we intend to use WordsEye itself to allow users to visualize vignettes as they define them, as a way to improve vignette accuracy and relevancy to the actual use of the system.

While the population of VigNet is not the focus of this paper, it is our goal to create a usable resource that can be populated with a reasonable amount of effort. We note that opposed to resources like FrameNet that require skilled lexicographers, we only need simple visual annotation that can easily be done by untrained Mechanical Turkers. In addition, as described in section 3.2, vignettes defined at more abstract levels of the frame hierarchy can be used and composed to cover large numbers of frames in a plausible manner. This allows more specific vignettes to be defined where the differences are most significant. VigNet is is focused on visually-oriented language involving tangible objects. However, abstract, process-oriented language and relations such as negation can be depicted iconically with general vignettes. Examples of these can be seen in the figurative and metaphorical depictions shown in (Coyne and Sproat, 2001).

5.2 Using VigNet in Text-to-Scene Generation

To compose a scene from text input such as *the man is washing the apple* it is necessary to parse the sentence into a semantic representation (evoking frames for each content word) and to then resolve the language-level semantics to a set of graphical entities and relations. To create a low-level graphical representation all frame elements need to be filled with appropriate semantic nodes. Frames support the selection of these nodes by specifying constraints on them using selectional restrictions. The

SUBFRAME-PARALLEL decomposition of vignettes then ultimately relates these nodes using elementary spatial vignettes (FRONTOF, ON, ...).

Note that it is possible to describe scenes directly using these vignettes (such as *The man is in front of the sink. He is holding an apple.*), as was used to create the mock-ups in figure 1.

Vignettes can be directly applied or composed together. Composing vignettes involves unifying their frame elements. For example, in washing an apple, the WASH-SMALL-FRUIT vignette uses a sink. From world knowledge we know (via instances of the TYPICAL-LOCATION frame) that washing food typically takes place in the KITCHEN. To create a scene we compose the two vignettes together by unifying the sink in the location vignette with the sink in the action vignette.

6 Related Work

The grounding of natural language to graphical relations has been investigated in very early text-to-scene systems (Boberg, 1972), (Simmons, 1975), (Kahn, 1979), (Adorni et al., 1984), and then later in Put (Clay and Wilhelms, 1996), and WordsEye (Coyne and Sproat, 2001). Other systems, such as CarSim (Dupuy et al., 2001), Jack (Badler et al., 1998), and CONFUCIUS (Ma and McKevitt, 2006) target animation and virtual environments rather than scene construction. A graphically grounded lexical-semantic resource such as VigNet would be of use to these and related domains. The concept of vignettes as graphical realizations of more general frames was introduced in (Coyne et al., 2010).

In addition to FrameNet, much work has been done in developing theories and resources for lexical semantics and common-sense knowledge. VerbNet (Kipper et al., 2000) focuses on verb subcat patterns grouped by Levin verb classes (Levin, 1993), but also grounds verb semantics into a small number of causal primitives representing temporal constraints tied to causality and state changes. VerbNet lacks the ability to compose semantic constraints or use arbitrary semantic relations in those constraints. Conceptual Dependency theory (Schank and Abelson, 1977) specifies a small number of state-change primitives into which all verbs are reduced. Event Logic (Siskind, 1995) decomposes actions into intervals describing state changes and allows visual grounding by specifying truth conditions for a small set of spatial primitives (a similar formalism is used by Ma and McKevitt (2006)). (Bailey et al., 1998) and related work proposes a representation in many ways similar to ours, in which lexical items are paired with a detailed specification of actions in terms of elementary body poses and movements. In contrast to these temporally-oriented approaches, VigNet grounds semantics in spatial constraints active at a single moment in time. This allows for and emphasizes *contextual reasoning* rather than causal reasoning. In addition, VigNet emphasizes a holistic frame semantic perspective, rather than emphasizing decomposition alone. Several resources for common-sense knowledge exist or have been proposed. In OpenMind and ConceptNet (Havasi et al., 2007) online crowd-sourcing is used to collect a large set of common-sense assertions. These assertions are normalized into a set of a couple dozen relations. The Cyc project is using the web to augment its large ontology and knowledge base of common sense knowledge (Matuszek et al., 2005). PRAXICON (Pastra, 2008) is a grounded conceptual resources that integrates motor-sensoric, visual, pragmatic and lexical knowledge (via WordNet). It targets the embodied robotics community and does not directly focus on scene generation. It also focuses on individual lexical items, while VigNet, like FrameNet, takes syntactic context into account.

7 Conclusion

We have described a new semantic paradigm that we call vignette semantics. Vignettes are extensions of FrameNet frames and represent the specific ways in which semantic frames can be realized in the world. Mapping frames to vignettes involves translating between high-level frame semantics and the lower-level relations used to compose a scene. Knowledge about objects, both in terms of their semantic types and the affordances they provide is used to make that translation. FrameNet frames, coupled with semantic nodes representing entity classes, provide a powerful relational framework to express such knowledge. We are developing a new resource VigNet which will implement this framework and be used in our text-to-scene generation system.

References

G. Adorni, M. Di Manzo, and F. Giunchiglia. 1984. Natural Language Driven Image Generation. In *Proceedings of COLING 1984*, pages 495–500, Stanford, CA.

N. Badler, R. Bindiganavale, J. Bourne, M. Palmer, J. Shi, and W. Schule. 1998. A parameterized action representation for virtual human agents. In *Workshop on Embodied Conversational Characters*, Tahoe City, CA.

D. Bailey, N. Chang, J. Feldman, and S. Narayanan. 1998. Extending Embodied Lexical Development. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Madison, WI.

C. Baker, C. Fillmore, and J. Lowe. 1998. The Berkeley Framenet Project. In *Proceedings of COLING 1998*, pages 86–90.

R. Boberg. 1972. Generating line drawings from abstract scene descriptions. Master's thesis, Dept. of Elec. Eng, MIT, Cambridge, MA.

S. R. Clay and J. Wilhelms. 1996. Put: Language-based interactive manipulation of objects. *IEEE Computer Graphics and Applications*, 16(2):31–39.

B. Coyne and R. Sproat. 2001. WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the Annual Conference on Computer Graphics*, pages 487–496, Los Angeles, CA.

B. Coyne, O. Rambow, J. Hirschberg, and R. Sproat. 2010. Frame Semantics in Text-to-Scene Generation. In *Proceedings of the KES'10 workshop on 3D Visualisation of Natural Language*, Cardiff, Wales.

S. Dupuy, A. Egges, V. Legendre, and P. Nugues. 2001. Generating a 3D Simulation Of a Car Accident from a Written Description in Natural Language: The CarSim System. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pages 1–8, Toulouse, France.

C. J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul.

C. Havasi, R. Speer, and J. Alonso. 2007. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Proceedings of RANLP 2007*, Borovets, Bulgaria.

K. Kahn. 1979. *Creation of Computer Animation from Story Descriptions*. Ph.D. thesis, MIT, AI Lab, Cambridge, MA.

H. Kamp. 1981. A Theory of Truth and Semantic Representation. In Groenendijk, J. and Janssen, T. and Stokhof, M., editor, *Formal Methods in the Study of Language*, pages 277–322. de Gruyter, Amsterdam.

K. Kipper, H. T. Dang, and M. Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of AAAI 2000*, Austin, TX.

B. Levin. 1993. *English verb classes and alternations: a preliminary investigation*. University Of Chicago Press.

M. Ma and P. McKevitt. 2006. Virtual human animation in natural language visualisation. *Artificial Intelligence Review*, 25:37–53, April.

C. Matuszek, M. Witbrock, R. C. Kahlert, J. Cabral, D. Schneider, P. Shah, and D. Lenat. 2005. Searching for Common Sense: Populating Cyc from the Web. In *Proceedings of AAAI 2005*, pages 1430–1435, Pittsburgh, PA.

K. Pastra. 2008. PRAXICON: The Development of a Grounding Resource. In *Proceedings of the International Workshop on Human-Computer Conversation*, Bellagio, Italy.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. R. Johnson, and J. Scheffczyk. 2010. *Framenet II: Extended Theory and Practice*. ICSI Berkeley.

R. C. Schank and R. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Earlbaum, Hillsdale, NJ.

R. Simmons. 1975. The CLOWNS Microworld. In *Proceedings of the Workshop on Theoretical Issues in Natural Language Processing*, pages 17–19, Cambridge, MA.

J. M. Siskind. 1995. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391.

R. Sproat. 2001. Inferring the environment in a text-to-scene conversion system. In *International Conference on Knowledge Capture*, Victoria, BC.