# The Development of Multimodal Lexical Resources

**James Pustejovsky, Nikhil Krishnaswamy, Tuan Do, Gitit Kehat**
Department of Computer Science
Brandeis University
Waltham, MA 02453 USA
{jamesp,nkrishna,tuandn,gititkeh}@brandeis.edu

## Abstract

Human communication is a multimodal activity, involving not only speech and written expressions, but intonation, images, gestures, visual clues, and the interpretation of actions through perception. In this paper, we describe the design of a multimodal lexicon that is able to accommodate the diverse modalities that present themselves in NLP applications. We have been developing a multimodal semantic representation, VoxML, that integrates the encoding of semantic, visual, gestural, and action-based features associated with linguistic expressions.

## 1 Motivation and Introduction

The primary focus of lexical resource development in computational linguistics has traditionally been on the syntactic and semantic encoding of word forms for monolingual and multilingual language applications. Recently, however, several factors have motivated researchers to look more closely at the relationship between both spoken and written language and the expression of meaning through other modalities. Specifically, there are at least three areas of CL research that have emerged as requiring significant cross-modal or multimodal lexical resource support. These are:

- **Language visualization and simulation generation**: Creating images from linguistic input; generating dynamic narratives in simulation environments from action-oriented expressions;(Chang et al., 2015; Coyne and Sproat, 2001; Siskind, 2001; Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016)
- **Visual Question-Answering and image content interpretation**: QA and querying over image datasets, based on the vectors associated with the image, but trained on caption-image pairings in the data; (Antol et al., 2015; Chao et al., 2015a; Chao et al., 2015b)
- **Gesture interpretation**: Understanding integrated spoken language with human or avatar-generated gestures; generating gesture in dialogue to supplement linguistic expressions;(Rautaray and Agrawal, 2015; Jacko, 2012; Turk, 2014; Bunt et al., 1998)

To meet the demands for a lexical resource that can help drive such diverse applications, we have been pursuing a new approach to modeling the semantics of natural language, *Multimodal Semantic Simulations (MSS)*. This framework assumes both a richer formal model of events and their participants, as well as a modeling language for constructing 3D visualizations of objects and events denoted by natural language expressions. The Dynamic Event Model (DEM) encodes events as programs in a dynamic logic with an operational semantics, while the language VoxML, Visual Object Concept Modeling Language, is being used as the platform for multimodal semantic simulations in the context of human-computer communication, as well as for image- and video-related content-based querying.

Prior work in visualization from natural language has largely focused on object placement and orientation in static scenes (Coyne and Sproat, 2001; Siskind, 2001; Chang et al., 2015). In previous work (Pustejovsky and Krishnaswamy, 2014; Pustejovsky, 2013a), we introduced a method for modeling natural language expressions within a 3D simulation environment, Unity. The goal of that work was to



*Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces*,
pages 41–47, Osaka, Japan, December 11 2016.

evaluate, through explicit visualizations of linguistic input, the semantic presuppositions inherent in the different lexical choices of an utterance. This work led to two additional lines of research: an explicit encoding for how an object is itself situated relative to its environment; and an operational characterization of how an object changes its location or how an agent acts on an object over time. The former has developed into a semantic notion of situational context, called a *habitat* (Pustejovsky, 2013a; McDonald and Pustejovsky, 2014), while the latter is addressed by dynamic interpretations of event structure (Pustejovsky and Moszkowicz, 2011b; Pustejovsky, 2013b; Mani and Pustejovsky, 2012; Pustejovsky, 2013a). The requirements on a "multimodal simulation semantics" include, but are not limited to, the following components:

- A minimal embedding space (MES) for the simulation must be determined. This is the 3D region within which the state is configured or the event unfolds;
- Object-based attributes for participants in a situation or event need to be specified; e.g., orientation, relative size, default position or pose, etc.;
- An epistemic condition on the object and event rendering, imposing an implicit point of view (POV);
- Agent-dependent embodiment; this determines the relative scaling of an agent and its event participants and their surroundings, as it engages in the environment.

In the sections that follow, we outline briefly the components of a multimodal lexical entry to address the needs stated above by the CL community for the first two areas. Integration of gesture interpretation and modeling is presently ongoing work in our lab.

## 2 VoxML: a Language for Concept Visualization

While both experience and world knowledge about objects and events can influence our behavior, as well as the interpretation and consequences of events, such factors are seldom involved in representing the predicative force of a particular lexeme in a language. Some representations, such as Qualia Structure (Pustejovsky, 1995) do provide additional information that can be used to map a linguistic expression to a minimal model of the event, and then from there to a visual output modality such as one that may be produced by a computer system, and so requires a computational framework to model it. Still, such languages are not in themselves rich enough to create useful minimal models.

To remedy this deficit, we have developed modeling language VoxML (Visual Object Concept Markup Language) for constructing 3D visualizations of natural language expressions (Pustejovsky and Krishnaswamy, 2016). VoxML forms the scaffold used to link lexemes to their visual instantiations, termed the "visual object concept" or *voxeme*. In parallel to a lexicon, a collection of voxemes is termed a *voxicon*. There is no requirement on a voxicon to have a one-to-one correspondence between its voxemes and the lexemes in the associated lexicon, which often results in a many-to-many correspondence. That is, the lexeme *plate* may be visualized as a [[SQUARE PLATE]], a [[ROUND PLATE]][1], or other voxemes, and those voxemes in turn may be linked to other lexemes such as *dish* or *saucer*.

Each voxeme is linked to an object geometry (if a noun—OBJECT in VoxML), a DITL program (if a verb or VoxML PROGRAM), an attribute set (VoxML ATTRIBUTEs), or a transformation algorithm (VoxML RELATIONs or FUNCTIONs). VoxML is used to specify the "episemantic" information beyond that which can be directly inferred from the geometry, DITL (Pustejovsky and Moszkowicz, 2011a), or attribute properties.

In order to demonstrate the composition of the linguistic expression plus the VoxML encoded information into a fully-realized visual output, we have developed, **VoxSim** (Krishnaswamy and Pustejovsky, 2016), a visual semantic simulator built on top of the Unity game engine (Goldstone, 2009).[2]

[1]Note on notation: discussion of voxemes in prose will be denoted in the style [[VOXEME]] and should be taken to refer to a visualization of the bracketed concept.

[2]The VoxSim Unity project and source may be found at `https://github.com/nkrishnaswamy/voxicon`.

VoxSim does not rely on manually-specified categories of objects with identifying language, and instead procedurally composes the properties of voxemes in parallel with the lexemes to which they are linked. Input is a simple natural language sentence, which is part-of-speech tagged, dependency-parsed, and transformed into a simple predicate-logic format.

From tagged and parsed input text, all NPs are indexed to objects in the scene. A reference to *a/the ball* causes the simulator to attempt to locate a voxeme instance in the scene whose lexical predicate is "ball," while an occurrence of *a/the block* prompts an attempt to locate a voxeme with the lexical predicate "block". Attributive adjectives impose a sortal scale on their heads, so *small block* and *big block* single out two separate blocks if they exist in the scene, and the VoxML-encoded semantics of "small" and "big" discriminates the blocks based on their relative size. *red block* vs. *green block* results in a distinction based on color, a nominal attribute, while *big red block* and *small red block* introduce scalar attribution, and can be used to disambiguate two distinct red blocks by iteratively evaluating each interior term of a formula such as *big(red(block))* until the reference can be resolved into a single object instance in the scene that has all the signaled attributes[3]. The system may ask for clarification (e.g., *"Which block?"*) if the object reference is still ambiguous.

An OBJECT voxeme's semantic structure provides *habitats*, situational contexts or environments which condition the object's *affordances*, which may be either "Gibsonian" and "telic" *affordances* (Gibson et al., 1982; Pustejovsky, 1995; Pustejovsky, 2013a). Affordances are used as attached behaviors, which the object either facilitates by its geometry (Gibsonian) (Gibson et al., 1982), or purposes for which it is intended to be used (telic) (Pustejovsky, 1995). For example, a Gibsonian affordance for [[CUP]] is "grasp," while a telic affordance is "drink from." Following from the convention that agents of a VoxML PROGRAM must be explicitly singled out in the associated implementation by belonging to certain entity classes (e.g., humans), affordances describe what *can be done to* the object, and not what actions it *itself* can perform. Thus an affordance is notated as HABITAT $\rightarrow$ [EVENT]RESULT, and an instance such as $H_{[2]} \rightarrow [put(x, on([1]))support([1], x)$ can be paraphrased as "In habitat-2, an object $x$ can be put on component-1, which results in component-1 supporting $x$." This procedural reasoning from habitats and affordances, executed in real time, allows VoxSim to infer the complete set of spatial relations between objects at each state and track changes in the shared context between human and computer. Thus, simulation becomes a way of tracing the consequences of linguistic spatial cues through a narrative.

A VoxML entity's interpretation at runtime depends on the other entities it is composed with. A cup on a surface, with its opening upward, may afford containing another object, so to place an object $in(cup)$, the system must first determine if the intended containing object (i.e., the cup) affords containment by default by examining its affordance structure.

If so, the object must be currently situated in a *habitat* which allows objects to be placed partially or completely inside it (represented by RCC relations PO, TPP, or NTPP). In VoxML, [[CUP]] is encoded as a concave object with rotational symmetry around the Y-axis and reflectional symmetry across the XY and YZ planes, meaning that it opens along the Y-axis. Its HABITAT further situates the opening along its positive Y-axis, meaning that if the cup's opening along its +Y is currently unobstructed, it affords containment. Previously established habitats, i.e., "The cup is flipped over," may activate or deactivate these and other affordances.

The spatial relations operating within the context of a verbal program, such as "put the spoon in the cup," enforce constraints that requires a test against the current situational context before a value assignment can be made. Given *put*, if the "placed object" is of a size that is too large to fit inside the mentioned object, VoxSim conducts a series of calculations to see if the object, when reoriented along any of its three orthogonal axes, will be situated in a configuration that allows it to fit inside the region bounded by the ground object's containing area. The containing area is situated relative to one of the ground object's orthogonal axes, and which axis and orientation this is is encoded in the ground object's VoxML type semantics. For example, the symmetrical and concave properties of [[CUP]] compose to situate the cup's opening along its *positive* Y-axis. So, to place a [[SPOON]] in a [[CUP]], assuming

[3] See (Pustejovsky and Krishnaswamy, forthcoming) for details on discriminating and referencing objects through sortal and scalar descriptions.

objects of typical size, [[SPOON]] must be reoriented so that its world-space bounding box aligning with the [[CUP]]'s Y-axis is smaller than the bounds of the [[CUP]]'s opening in that same configuration.

## 3 Video Event Recognition

Now let us turn the language-to-visualization strategy on its head. The same modeling language, VoxML, can be used to help detect and recognize events and actions in video. This task has received increasing attention in the scientific community, due to its relevance to a wide variety of applications (Ballan et al., 2011) and there have been calls for annotation infrastructure that includes video (Ide, 2013).

Our lab has begun bootstrapping a dataset of videos annotated with event-subevent relations using ECAT, an internally-developed video annotation tool (Do et al., 2016). This annotation tool allows us to annotate videos of labeled events with object participants and subevents, and to induce what the common subevent structures are for the labeled superevent. Using the Microsoft Kinect®, we are currently recording videos of a test set of human actions interacting with simple objects, such as blocks, cylinders, and balls. Both human bodies (rigs) and these objects can be tracked and annotated as participants in a recorded motion event; this labeled data can then be used to build a corpus of *multimodal semantic simulations* of these events that can model object-object, object-agent, and agent-agent interactions through the durations of said events. This library of simulated motion events can serve as a novel resource of direct linkages from natural language to event visualization, indexed through the multimodal lexical representation for the event, its voxeme.

We are also interested in leveraging VoxML PROGRAMs to facilitate machine learning algorithms in activity recognition. Our motivation is that modeling actions as a rigorous dynamic structures allows us to represent action as labelled state transition systems. Therefore, we can model their similarity and difference using classical graph similarity approaches. For example, we aim to reveal in the data the intuition that there is a similarity between "I toss a ball" and "I jump in the air", i.e. a *figure* object moving in the same manner in relative to *ground* object. This is different from other activity recognition approaches, such as (Shahroudy et al., 2016), in which the authors directly used supervision learning on different classes of activities.

We have begun creating lexical resources using movie databases, such as MPII Movie Description Dataset (Rohrbach et al., 2015), which has parallel movie snippets and descriptions. These descriptions are transcribed from audio descriptions for the visually impaired. Therefore, they are highly event-centric, describing the most salient events in each movie snippet. By annotating them using the same annotation framework as mentioned above for the 3D motion capture, we aim to create a rich word sense resource. In turn, we hope that we can use these modes of representation to discover the difference between *canonical* and *non-canonical* uses of activity verbs.

## 4 Image Grounding for the Lexicon

Finally, another aspect of multimodal lexicalized meaning that we are investigating, and which has become increasingly popular among both the computer vision and NLP communities, is the creation and usage of vision-language datasets. These datasets typically contain still images along with a set of textual annotations, such as nouns, attributes and verbs, or full descriptive sentences or Q&A pairs, for each image in the dataset. They are mostly used in the training and evaluation of tasks sitting at the intersection of vision and language, such as image description generation, visual question answering and image retrieval, but they are also used in tasks such as action and affordance recognition, to support and expand previous "vision-only" techniques with linguistic information.

As the field is growing, more time and effort are being spent on the creation of large-scale vision-language datasets (Krishna et al., 2016; Lin et al., 2014), as well as smaller task-oriented ones for tasks like the ones mentioned above (Chao et al., 2015a; Chao et al., 2015b). However, we found that many of the existing datasets suffer from problems making them difficult to use in a consistent way (Kehat and Pustejovsky, 2016). Some of the main difficulties are: vocabulary issues (both limited or sparse);

lack of validation or averaging process that leads to information loss; a heavy bias originated in both the authors pre-assumptions and annotators attentio; and underdefined visual actions/concepts. The last problem, which is perhaps the most challenging of all, is related to the fact that in the majority of datasets, verbs and actions are considered the same. However, in reality, one verb can describe multiple different visually defined actions, and the same visual action can be matched to more than one verb. While most of the existed datasets do not distinguish between the two, there are new attempts to solve this inherent ambiguity, as well as to define what a visually defined action is (Gella et al., 2016; Ronchi and Perona, 2015; Yatskar et al., 2016).

## 5 Conclusion and Future Directions

We have described our initial steps towards the design and development of a multimodal lexical resource, based on a modeling language that admits of multiple representations from different modalities. These are not just linked lists of modal expressions but are semantically integrated and interpreted representations from one modality to another. The language VoxML and the resource Voxicon are presently being used to drive simulations using multiple modalities within the DARPA Communicating with Computers program.

## Acknowledgements

This work is supported by a contract with the US Defense Advanced Research Projects Agency (DARPA), Contract W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We would like to thank Scott Friedman, David McDonald, Marc Verhagen, and Mark Burstein for their discussion and input on this topic. All errors and mistakes are, of course, the responsibilities of the authors.

## References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. 2011. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302.

Harry Bunt, Robbert-Jan Beun, and Tijn Borghuis. 1998. *Multimodal human-computer communication: systems, techniques, and experiments*, volume 1374. Springer Science & Business Media.

Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. 2015. Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015a. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025.

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015b. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4259–4267.

Bob Coyne and Richard Sproat. 2001. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496. ACM.

Tuan Do, Nikhil Krishnaswamy, and James Pustejovsky. 2016. Ecat: Event capture annotation tool. *Proceedings of ISA-12: International Workshop on Semantic Annotation*.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 182-192. San Diego.*

James Jerome Gibson, Edward S Reed, and Rebecca Jones. 1982. *Reasons for realism: Selected essays of James J. Gibson*. Lawrence Erlbaum Associates.

Will Goldstone. 2009. *Unity Game Development Essentials*. Packt Publishing Ltd.

Nancy Ide. 2013. An open linguistic infrastructure for annotated corpora. In *The People's Web Meets NLP*, pages 265–285. Springer.

Julie A Jacko. 2012. *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. CRC press.

Gitit Kehat and James Pustejovsky. 2016. Annotation methodologies for vision and language dataset creation. *IEEE CVPR Scene Understanding Workshop (SUNw), Las Vegas*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Nikhil Krishnaswamy and James Pustejovsky. 2016. Multimodal semantic simulations of linguistically underspecified motion events. *Proceedings of Spatial Cognition*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Inderjeet Mani and James Pustejovsky. 2012. *Interpreting Motion: Grounded Representations for Spatial Language*. Oxford University Press.

David McDonald and James Pustejovsky. 2014. On the representation of inferences and their lexicalization. In *Advances in Cognitive Systems*, volume 3.

James Pustejovsky and Nikhil Krishnaswamy. 2014. Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)*, page 99.

James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A visualization modeling language. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

James Pustejovsky and Nikhil Krishnaswamy. forthcoming. Envisioning language: The semantics of multimodal simulations.

James Pustejovsky and Jessica Moszkowicz. 2011a. The qualitative spatial dynamics of motion. *The Journal of Spatial Cognition and Computation*.

James Pustejovsky and Jessica L Moszkowicz. 2011b. The qualitative spatial dynamics of motion in language. *Spatial Cognition & Computation*, 11(1):15–44.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2013a. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10. ACL.

James Pustejovsky. 2013b. Where things happen: On the semantics of event localization. In *Proceedings of ISA-9: International Workshop on Semantic Annotation*.

Siddharth S Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description.

Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing common human visual actions in images. *arXiv preprint arXiv:1506.02203*.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *arXiv preprint arXiv:1604.02808*.

Jeffrey Mark Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.(JAIR)*, 15:31–90.

Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. *In Proceedings of the Conference of Computer Vision and Pattern Recognition (CVPR)*.