PRAXICON: The Development of a Grounding Resource

Katerina Pastra[1]

Abstract. In this paper, we trace the different manifestations of grounding resources in a variety of Artificial Intelligence applications and present the POETICON project's development plans for building such resource. In doing so, we introduce the *PRAXICON*, a resource that links natural language and sensorimotor representations of concepts, with the aim of facilitating multimodal and multimedia content integration in cognitive systems. We correlate the envisaged resource with semantic language resources and briefly discuss the possibility of interfacing these different types of resources.

1 Introduction

The computational integration of sensorimotor and symbolic representations (grounding) and in particular of visual (visual objects and actions) and natural language representations has a long history that could be traced back to the very early days of Artificial Intelligence (AI). In comparing language and vision and arguing on how the corresponding representations are associated, Jackendoff has *implied* the use of a kind of integrated, grounding resource in the human memory for use in both natural language and vision understanding [23, 24]. He suggested that a resource with both conceptual and visual information should be built and in particular, one that includes conceptual structures [22] at the leaf nodes of which concepts are linked to corresponding 3D models of visual objects (cf. the notion of 3D model catalogue in Marr's theory of vision [33]). In fact, the idea of building a resource that will link visual and linguistic representations goes back at least to the seventies [3]; it was a reduced version of Jackendoff's suggestion and involved the association of manually detected image regions (objects) and corresponding image feature vectors with words (names for the objects); a similar type of resource has been suggested recently under the term of *multimedia thesaurus* for use in various multimedia processing tasks [10, 5, 25, 54]. In fact, the idea for such resources has been lurking around for decades, never explicitly said or described and only partially implemented in an *ad hoc* basis in AI [43, 40].

In this paper, we introduce the notion of a *PRAXICON*, a resource that bridges natural language and sensorimotor representations of concepts, with the aim of facilitating multimodal and multimedia content integration in cognitive systems. In what follows, we present the "precursors" of grounding resources that have been developed for a wide range of AI applications and needs, and elaborate on the notion of the PRAXICON as suggested within the framework of the EC-funded research and development project, POETICON. We focus on the development plan and the envisaged contents of the PRAXICON and discuss the role *semantic language resources* could play in its development.

[1] Language Technologies Department, Institute for Language and Speech Processing, Athens, Greece email: kpastra@ilsp.gr

2 Grounding Resources

In practice, grounding resources have been used in AI prototypes that span a number of decades and a wide range of application areas, from the SHRDLU system [59], which verbalized visual changes in a 2D blocks scene, to *medium translation* systems (e.g. automatic sports commentators), *multimedia presentation* systems (e.g. automatic creation of illustrated technical manuals) and *conversational robots* of the new millennium (cf. a related state of the art review of more than 50 such prototypes in [43, 40]). In most of these prototypes, the grounding resources link words denoting entities with corresponding low-level visual features of objects, and/or the actual images of the objects (e.g. superquadric images or wireframe models or 2D drawing patterns etc.). The words are then linked to conceptual structures which take the forms of ontologies, e.g. [55], semantic frames or schemas, e.g. [34, 51, 8], object or scene domain models, e.g. [7]. Prototypes that involve dynamic scenes/motion rather than static objects, use resources that link words denoting motion with corresponding motion trajectory information; the words are then linked to conceptual structures denoting events, in the form of event templates, e.g. [11]. In other cases, manually or semi-automatically extracted geometric scene description data which link object information, their motion trajectory and their name/word is used [1, 20, 32]; this first grounding level is correlated to a hierarchy of event models which includes event concepts from general to specific that reflect temporal decomposition, e.g. [21]. In conversational robots, resources that link words with corresponding n-ary tuples of feature-values of visual objects and functions that implement motion have been used; the links are then associated to conceptual structures denoting classes of entities/events [49, 45].

In all these cases, the grounding resources used have the following characteristics:

1. they have been created ad hoc, for the needs of a very limited prototype that works either in a blocksworld (blobs) or miniworld (very limited and simplified domain); the objects and/or events included in the resource are very simple (e.g. soccer ball, motion of the ball along the playfield), the resources themselves have not been reused in any way;
2. there is no exploration of the association level at which words and sensorimotor representations are connected, and no systematic study on how different representations collaborate in forming higher-level concepts; it is intuition rather than a rigorous methodology that lies behind their creation;
3. the visual and motoric representations used in the resources are manually created, i.e. visual abstraction is not done automatically, with the exception of some work on automatic 3D trajectory extraction (used in a verbalization task of traffic scenes, [16, 2]), which is still limiting motion representations to the motion path of a rigid object (as opposed to the fine-grained movements of

a human body); sensorimotor representations are monolithic/flat feature-value representations.

Scaling up such resources using automatic mechanisms for associating sensorimotor and symbolic representations and enhancing their coverage is highly questionable. In another strand of research that has been using *language games* as a method to teach agents primitive models of language [53], or how to associate words with objects [26], machine learning methods have been followed. Instead of using a manually created, fixed grounding resource, the agent is being endowed with a learning module which allows it to develop such resource on its own through e.g. supervised learning techniques. A human feeds the robot with a word which corresponds to the image of an object in view; the robot is able to generate a simplified representation of the image of this object in the form of feature-value vectors standing for the colour and shape attributes of the object. Once a number of associations are learned, the agent compares the (representation of the) image of any new object with the ones it knows using e.g. a nearest neighbour algorithm and associates the new object with the name of the known object it is more similar to. Similar statistical or connectionist algorithms for associating words in an utterance with objects in view have also been developed in a limited number of artificial agents [48, 39, 46, 57, 44, 6].

This approach in developing a grounding resource has most of the limitations of the symbolic approaches mentioned above: the resource built by the agent is limited to blobs or simple objects (motoric representations are not treated at all), it relies on the developer's intuition for choosing the words to be associated with the visual representation of an object, while the developer intervenes in visual abstraction by presenting to the agent a clean image of an object (i.e. one stripped from any background, or the presence of other objects which could lead to wrong associations, etc.). More importantly, the resource does not connect the words the agent learns with higher-level conceptual structures (i.e. more complex concepts), which renders scaling for the needs of everyday interaction questionable.

In the last few years, image-language association mechanisms as such are being developed for automatic image/keyframe annotation, with the vision of being, at some point, mature enough for being embedded in multimedia prototypes and mainly in *multimedia indexing and retrieval* prototypes. The approaches are either probabilistic [4, 58] or logic-based [9]. Learning approaches rely on properly annotated training corpora for learning the associations between images/image regions represented in feature-value vectors and corresponding textual labels, cf. for example IBM's corpus [30] and the PASCAL visual object categorization collection [12]. Symbolic logic approaches rely on *multimedia ontologies*, i.e. hierarchical conceptual resources the leaf nodes of which (words) are associated with corresponding image feature-vectors [9, 50] or with a set of corresponding actual images collected from the web [60]. Srikanth et al. [52] report also on the use of both training corpora and ontologies for achieving automatic image annotation. The annotated training corpora could be thought of as simplistic grounding resources (with no connection to conceptual structures), while the ontologies do provide access to higher-level concepts; however all above mentioned criticisms apply to both of them (i.e. limited and non-systematic development, association based on intuition, problems in scaling up etc.).

In spanning such wide range of applications, these preliminary "grounding resources" point to the necessity of investing research effort towards their development. However, how should one proceed in building such resource, if one is to avoid pitfalls and limitations of previous attempts and more significantly, what is it exactly that such resource should comprise of?

3 PRAXICON: the POETICON Suggestion

Developing an automatically extensible grounding resource is the main objective of POETICON, a newly funded EC-FP7 research project (http://www.poeticon.eu). POETICON explores *the poetics of everyday life*, i.e. the synthesis of sensorimotor representations and natural language in everyday human interaction. As shown in the previous section, this is an old problem, one that relates to the long Artificial Intelligence (AI) quest for meaning.

However, while meaning extraction and generation with sensorimotor or symbolic representations (i.e. moving or static images, action, text/speech) has been the objective in most AI sub-disciplines (e.g. Natural Language Processing, Image Understanding etc.) the research focus has mostly been on individual types of representations (e.g. visual, motoric or linguistic) rather than on their integration [43, 40]. There are a number of interrelated reasons that have hindered research to focus on such integration: *difficulties in measuring sensorimotor human behaviour, and analyzing visual and motoric representations*, *the tendency for isolation among researchers* working on e.g. extracting meaning from language or images, *lack of a corresponding theoretical, cognitive or/and neurophysiological background* that would provide a solid basis for computational investigation of the issue.

POETICON suggests a new approach within this AI quest for meaning; it *captures/measures* and *computationally structures* human behaviour (human body movements, facial expressions, manipulation of objects) with the objective of creating a **PRAXICON**, a grounding resource, and mechanisms that will facilitate its automatic extension. In particular, the PRAXICON will be the result of the following:

The POETICON corpus: a corpus of four distinct though interrelated- sets of multisensory recordings of human movements, visual objects, facial expressions and *enacted everyday scenarios*. In the latter, human body movements, facial expressions, objects and natural language interact in forming meaning in human to human interaction. The corpus supports analysis across a large number of dimensions from 2D analysis of video streams to complex models of 3D articulation.

The Human Activity Language (HAL) [17]: a hierarchical structural analysis of motoric representations, i.e. an innovative computational analysis of action into primitive units and production rules for formulating more or less complex actions. The HAL parser for motoric representations is tightly coupled to a corresponding visual action parser. Both tools are used for generating motoric and visual representations of action-denoting concepts in the PRAXICON.

The Language of Facial Expressions [38]: a hierarchical analysis of the perception of facial expressions at great detail providing information about the vocabulary of facial expressions, i.e., the importance of individual facial regions for recognition as well as other tasks, and the grammar of facial expressions, i.e., the inter-regional timing as well as the spatial integration of facial regions how the individual regions have to be integrated both in space and time in order to create a facial expression that carries meaning. The analysis is used for the sensorimotor representation of affective concepts in the PRAXICON.

The Language of Visual Object Representations [15, 14] : a structural analysis of visual representations into primitive units and production rules for large-scale representation of visual object categories. A hierarchical representation of visual input is developed

that enables the recognition and detection of a large number of object categories. Inspired by the principles of efficient indexing, robust matching, and ideas of compositionality, the approach is to learn a hierarchy of spatially flexible compositions, i.e. parts, in an unsupervised, statistics-driven manner. The existence of a grammatical structure in the learned hierarchical visual lexicon that would alleviate the problem of large-scale representation of object categories is also investigated. The approach is used for the visual representations of object-denoting concepts in the PRAXICON.

Cognitive Experiments: human subjects are asked to describe verbally the contents of the videos of the POETICON everyday scenarios. In scaling the visual context of the actions/entities to be verbalized by the subjects, associations of sensorimotor and symbolic representations at various levels of abstraction are elicited. The descriptions vary from cognitive categorization of the smallest meaningful units of visual action/entities to categorization of whole scenes and to free story-telling commentary. The experiments will provide the link between natural language and sensorimotor representations that will be covered in the resource. This is a "black-box" exploration of how one talks about what one does and what one sees. A "glass-box" approach looking inside one's brain under the same conditions takes place through neurophysiological experiments.

Neurophysiological experiments: The experiments explore how motor synergies, thought to be in common between motor and linguistic domains, develop and how syntactically meaningful chains of movements are organized to achieve an action goal [13]. A hypothesis on the role of Broca's area is put forward; because of its premotor nature, the area was firstly involved in generating/extracting action meanings, and then this ability might have been generalized during evolution giving to this area the basics to build a new capability: a *supramodal syntax* endowed with the ability to organize and comprehend hierarchical and sequential elements into meaningful structures. The role of Broca's area is explored through behavioural experiments, both in normal humans and in frontal aphasics.

Grounding resource use and extensibility: a number of experiments that explore the extent to which the PRAXICON could be used in audiovisual data processing for associating visual action and visual object representations with natural language and how the resource could be expanded automatically are under way. The grounding resource is employed to identify associations between sensorimotor representations and language within an audiovisual file; the main challenge in this case is that the language that is used in conjunction with the visual representations of objects and movements (video footage) does not necessarily refer to what is depicted in the file. What one hears (speech) may refer to what one sees (at various levels of abstraction), but it might also be the case that what one sees and what one hears bring complementary or even contradictory information. We explore cases in which the word(s) used for the automatically perceived sensorimotor representations will not be the ones expected according to the resource (e.g. cases of synonymy, metaphor, antithesis). Also cases in which the sensorimotor representations perceived are not covered in the resource, though the corresponding word/concept is in the resource (case of polysemy) will be explored. In both cases, proper extension of the resource will take place automatically. Data mining experiments searching for structure between language and the underlying sensorimotor volume will take place. Relying on the output of the computational tools mentioned above (visual action and visual object parsers) that will run on this volume of data, we expect to find correlations between verbs and sensorimotor patterns. In other words, abstract verbs/nouns tell a story, and it will be that story captured in the sensorimotor correlates. How abstract the story is (i.e. how closely related to what is depicted) will be investigated based on the COSMOROE cross-media interaction relations framework and the corresponding annotated corpora [42].

Experimentation with a humanoid: The iCub humanoid platform [56] is extended with a tactile skin for reaching, grasping and manipulation activities, with integration abilities of proprioceptive and visual information for learning and recognition of multi-sensory object representations, and with well-grounded, empirical representations of perception and action. Using the tools and the grounding resource developed in the project, the robot will carry out behaviour analogous to the components of everyday tasks studied in humans. It is expected that the lowest-level details of such actions (that depend on the precise morphology of the robot) will require robot-specific representations, while the coarsest-level structure should be shared with representations recovered from human action (such as the overall sequencing of parts of the action). By learning where the boundary lies between the two, and how to manage their interface, we aim at generating a reusable procedure for applying the work of this project to robotic applications.

The resource will be unique in that it relies on the development of innovative computational tools that analyse sensorimotor representations of everyday activities, the use of cognitive methods for establishing the associations of such representations with natural language and the development of a learning mechanism for extending the resource automatically. Furthermore, neurophysiological experiments and experimentation with a humanoid are the driving forces and implementation tools respectively for the development of the resource. However, what are specifically the contents of such resource?

3.1 Elaborating on the notion of the PRAXICON

The PRAXICON is envisaged to be a computational resource which associates symbolic representations (words/concepts) with corresponding sensorimotor representations, and patterns of their combinations that formulate conceptual structures at different levels of abstraction. The resource is envisaged to allow artificial agents/systems:

- to tie concepts/words of different levels of abstraction to their sensorimotor instantiations (catering thus for disambiguation), and
- to untie sensorimotor representations from their physical specificities correlating them to conceptual structures of different levels of abstraction (catering thus for intentionality indication).

In other words, going bottom-up in the resource (from sensorimotor representations to concepts) one will get a hierarchical composition of human behaviour, while going top-down (from concepts to sensorimotor representations) one will get intentionality-laden interpretations of those structures.

The PRAXICON relies on the theoretical premise that meaning emerges also from the integration of sensorimotor and symbolic representations and in particular, it emerges from:

- the integration of different types of representations that refer to the same entity, event or property[2];
- the integration of representations that refer to different entities, events, or properties but collaborate in forming concepts at different levels of abstraction[3] [42].

[2] Cf. for example, the word put and the corresponding motoric representation.

[3] Cf. for example the following case: dress (word and visual representation) + put (word and motoric representation) + washing-machine (word and visual representation) to denote the concept of washing ones clothes.

This thesis is closely related to the Symbol Grounding theory [18, 19] which argues that grounding of symbols to sensorimotor experiences is necessary for AI agents to grasp the meaning of symbolic representations (natural language) and to respond appropriately. However, POETICON's thesis on meaning goes beyond traditional grounding approaches in that [43, 40]:

1. It considers grounding to be a bi-directional process (double-grounding), during which symbols are grounded to corresponding sensorimotor representations for getting tied to the physical entities/events/properties they refer to, and they also ground in their turn- the sensorimotor representations for enriching them with intentionality indicators. In other words, different aspects of meaning emerge through this two-way integration of symbolic and sensorimotor representations: meaning that disambiguates/clarifies linguistic reference (e.g. word-sense disambiguation), and meaning that disambiguates/clarifies sensorimotor reference (e.g. focus, salience, type-token distinctions), i.e. it renders intentionality in human behaviour (sensorimotor representations) explicit.
2. It does not consider all symbols able or suitable for grounding; some symbols do not need to be grounded, they are not meant to be tied to sensorimotor experience, they serve different purposes; i.e. to abstract away from the sensorimotor nature of human behaviour and comment on the functional, purposive, intentional nature of such behaviour[4].

3.2 Turning to Semantic Language Resources

In trying to pin down the notion of the PRAXICON, one could describe it drawing an analogy to lexicons:

- a *lexicon with grounded lemmas*, i.e. entries that comprise of a word sense and the corresponding visual and/or motoric representation
- a *lexicon with conceptual/pragmatic relations between grounded lemmas*, i.e. entries that are related with each other in different ways, forming conceptual structures of different levels of abstraction (e.g. action-object combinations, action-action combinations etc.)

However, in designing the PRAXICON one needs to go into detail on the nature of its contents and their optimal organisation. So, what will PRAXICON lemmas be like? Will everything be organised around the language-representation of each entry or around the sensorymotor part of it? Will the entry be a one-word or multi-word token? How specific or general will each entry be? What kind of conceptual relations will actually be captured in the resource? Are other types of information on concepts needed (e.g. grammatical, syntactic information on words etc.)?

In trying to address these questions, one would normally look into other grounding resources to explore how similar issues have been addressed in them. However, as argued in section 2, grounding resources have never been developed systematically, i.e. there is no grounding resource of any scale beyond ad hoc system development. Hardly do these ad hoc resources go beyond a first mapping of a word with a visual representation to conceptual structures of any complexity, and if they do (e.g. in the case of multimedia ontologies) they just match the grounded concepts to a pre-existing language-based representation of conceptual structures. Scaling, extending or even basic questions related to the design and development of a grounding resource have never been posed. One needs, therefore, to look into detail on the nature of a different type of resources which could shed light on the development of the PRAXICON and even contribute to it by providing parts of the content to be included/covered in the resource: *language resources that capture semantic information*.

Within the framework of POETICON, we carried out a comparative study of 19 such resources that span many decades and a variety of conceptual information contents [41] . We, therefore, re-visited a wide range of language resources that go beyond computational semantic lexica such as SIMPLE [29], to lexical resources such as WordNet [35], VerbNet [27] and FrameNet [47], to common-sense knowledge-bases such as CYC [28], ConceptNet [31] and Thought Treasure [37], just to name a few. The criterion for including a resource in the study was whether it went beyond morphological and syntactic relations to lexical semantic relations (e.g. synonymy, antonymy) and to more complex conceptual relations (e.g. temporal inclusion, manner etc.) that require one to cross over lexicalization into conceptual structures. The aim was to explore whether any of these resources could:

- contribute to the development of the PRAXICON, or
- get interfaced at some point (conceptual level) with the resource, enriching it, or
- be itself tuned/enhanced to qualify for playing the role of a PRAXICON.

The review is unique not only in that it spans resources developed in different AI sub-disciplines, with different methodologies and perspectives, for different applications, but mainly because it looks at these resources from a totally new perspective, which is how they could contribute to the need of developing grounding resources and mechanisms for artificial agents/systems.

After a thorough analysis of the resources in terms of their profiling, content and methodology followed for their development, a number of trends in language resources seem to emerge:

- most resources get extended so that they cover more types of conceptual information/relations, going also down to the level of specific instances and facts
- the resources get mapped to each other for greater usability
- there is a constant search for automatic or semi-automatic mechanisms for extending the coverage of the resources, and
- there is a growing development of the resources in different languages all mapped to each other

One of the main issues that seem to emerge in the construction of language resources is how to express meaning in a universal way, going beyond the language-specific expression of such meaning evidenced in grammatical and syntactic/semantic information.

The review has pointed to the fact that through AI history and beyond, researchers were trying to find the primitive concepts or features (feature bundles) to describe the world in a universal, language-independent way. They used categorization to organise and name everyday-experiences, concrete and abstract/mental entities, actions and properties. They also used story-telling like constructions to represent everyday-life scenarios and events.

However, the link to what is truly universal, our sensorimotor experiences, is missing. We envisage the PRAXICON to be a

[4] Consider for example the word shopping and the sensorimotor representations that form this everyday activity e.g. put (motoric representation) products (visual object representation) in a trolley (visual object representation), go to the till, pay and so on. A great number of sensorimotor representations which can be named with specific symbols/concepts/words are abstracted under the shopping concept; which also has a number of realizations (e.g. shopping for clothes, shopping at the super-market etc.). Cf. for example the notion of frames [36].

sensorimotor-centric (rather than word/language-centric) resource of conceptual structures. Based on the lessons we learn from the development of language resources and making use of the different kinds of conceptual information that one can draw from each resource, PRAXICON will be an attempt to built a grounding resource. In this resource, sensorimotor representations will be used to capture the common experiences, the specifics of everyday interaction, and language will be used to express how people abstract from such specifics, interpreting everyday experiences and drawing inferences from them.

Developing a **sensorimotor-centric** (rather than lemma/word centric) resource requires:

- a very different organisation/structuring than what one finds in existing language-resources, and
- concept analysis (decomposition) at a much finer-grained level than what is currently available.

In particular, it requires one to analyse a concept down to a level at which inferences, common-sense knowledge or background knowledge will be minimal (if not extinct). One needs to identify the conceptual level of analysis at which humans start making inferences on what an object/action is. The PRAXICON lower-level entries will not represent concepts that are readily lexicalised, but the ones at the level of which inferences come into play. The POETICON cognitive experiments on naming will not simply guide the association of words and sensorimotor representations in the PRAXICON; they will attempt to identify this conceptual entry-point, at which sensorimotor and symbolic integration takes place. The wealth of conceptual information captured in semantic language resources will be not only useful, but essential for enriching the PRAXICON and for developing mechanisms for automating its extension.

4 Conclusion

In this paper, we introduced the PRAXICON, a grounding resource that goes beyond symbolic representations of concepts to sensorimotor ones and beyond lexicalization to conceptual structures. We presented the POETICON project's approach for building such resource, and correlated the resource to semantic language resources. The PRAXICON is a suggestion for approaching meaning representation from a multimodal rather than language-only perspective, bringing semantic language resources closer to cognitive system development.

ACKNOWLEDGEMENTS

The research reported in the paper is being funded by the European Commission Framework Program Seven (FP7-ICT-215843).

REFERENCES

[1] E. André, G. Herzog, and T. Rist, 'On the simultaneous interpretation of real world image sequences and their natural language description: the system soccer', in *Proceedings of the European Conference on Artificial Intelligence*, pp. 449–454, (1988).

[2] M. Arens and H. Nagel, 'Behavioral knowledge representation for the understanding and creation of video sequences', in *Proceedings of the German Conference on Artificial Intelligence*, volume 2821 of *Lecture Notes in Artificial Intelligence*, pp. 149–163, (2003).

[3] R. Bajcsy and A. Joshi, 'The problem of naming shapes: Vision-language interface', in *Proceedings of the Theoretical Issues in Natural Language Processing*, pp. 157–161, (1978).

[4] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, 'Matching words and pictures', *Journal of Machine Learning Research*, **3**, 1107–1135, (2003).

[5] A. Benitez, J. Smith, and S. Chang, 'Medianet: A multimedia information network for knowledge representation', in *Proceedings of the IEEE SPIE International Conference on Multimedia Management Systems*, volume 4210, (2000).

[6] N. Bredeche, Y. Chevaleyre, J. Zucker, A. Drogoul, and G. Sabah, 'A meta-learning approach to ground symbols from visual percepts', *Robotics and Autonomous Systems*, **43**, 149–162, (2003).

[7] C. Callaway, B. Daniel, and J. Lester, 'Multilingual natural language generation for 3d learning environments', in *Proceedings of the Argentine Symposium on Artificial Intelligence*, pp. 177–190, (1999).

[8] B. Coyne and R. Sproat, 'Wordseye: An automatic text to scene conversion system', in *Proceedings of the International Conference on Computer Graphics and Interactive Technologies*, pp. 487–496, (2001).

[9] S. Dasiopoulou, V. Papastathis, V. Mezaris, I. Kompatsiaris, and M. Strintzis, 'An ontology framework for knowledge-assisted semantic video analysis and annotation', in *Proceedings of the International Workshop on Knowledge Markup and Semantic Annotation*, (2004).

[10] M. Dobie, R. Tansley, D. Joyce, M. Weal, P. Lewis, and W. Hall, 'A flexible architecture for content and concept based multimedia information exploration', in *Proceedings of the Challenge of Image Retrieval Conference*, pp. 1–12, (1999).

[11] S. Dupuy, A. Egges, V. Legendre, and P. Nugues, 'Generating a 3d simulation of a car accident from a written description in natural language: the carsim system', in *Proceedings of the Association of Computational Linguistics Workshop on Temporal and Spatial Reasoning*, pp. 1–8, (2001).

[12] M. Everingham, L. Van Gool, C. Williams, and A. Zisserman. Pascal visual object classes challenge results. World Wide Web (http://www.pascal-network.org/challenges/VOC/voc), 2005.

[13] L. Fadiga and L. Craighero, 'Cues on the origin of language. from electrophysiological data on mirror neurons and motor representations.', in *On Being Moved: From mirror neurons to empathy.*, ed., S. Braten, John Benjamins, (2007).

[14] S. Fidler, M. Boden, and A. Leonardis, 'Similarity-based cross-layered hierarchical representation for object categorisation', in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (2008).

[15] S. Fidler and A. Leonardis, 'Towards scalable representations of object categories: Learning a hierarchy of parts', in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (2007).

[16] R. Gerber and H. Nagel, '(mis?)-using drt for generation of natural language text from image sequences', in *Proceedings of the European Conference on Computer Vision*, volume 1407 of *Lecture Notes in Computer Science*, pp. 255–270, (1998).

[17] G. Guerra-Filho and Y. Aloimonos, 'A language for human action', *IEEE Computer Magazine*, **40**(5), 60–69, (2007).

[18] S. Harnad, 'Minds, machines and searle', *Journal of Theoretical and Experimental Artificial Intelligence*, **1**, 5–25, (1989).

[19] S. Harnad, 'The symbol grounding problem', *Physica D*, **42**, 335–346, (1990).

[20] G. Herzog, 'From visual input to verbal output in the visual translator', in *Proceedings of the American Association of Artificial Intelligence Fall Symposium on Computational Models for Integrating Language and Vision*, (1995).

[21] G. Herzog, C. Sung, E. André, W. Enkelmann, H. Nagel, T. Rist, W. Wahlster, and G. Zimmermann, 'Incremental natural language description of dynamic imagery', in *Proceedings of the International Conference on Artificial Intelligence*, pp. 153–162, (1989).

[22] R. Jackendoff, *Semantics and Cognition*, MIT Press, 1983.

[23] R. Jackendoff, 'On beyond zebra: the relation of linguistic and visual information', *Cognition*, **20**, 89–114, (1987).

[24] R. Jackendoff, *Languages of the mind: essays on mental representation*, MIT Press, 1992.

[25] D. Joyce, P. Lewis, R. Tansley, M. Dobie, and W. Hall, 'Semiotics and agents for integrating and navigating through multimedia representations of concepts', in *Proceedings of the International Conference on Storage and Retrieval for Media Databases*, volume 3972 of *SPIE Proceedings*, pp. 132–143, (2000).

[26] F. Kaplan, 'Talking aibo: First experimentation of verbal interactions with an autonomous four-legged robot', in *Proceedings of the TWENTE Workshop on Language Technology*, pp. 57–63, (2000).

[27] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, 'A large-scale clas-

sification of english verbs', *Language Resources and Evaluation*, **42**, 21–40, (2008).
[28] D. Lenat, 'Cyc: A large-scale investment in knowledge infrastructure', *Communications of the ACM*, **38**(11), 33–38, (1995).
[29] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli, 'Simple: A general framework for the develoment of multilingual lexicons', *International Journal of Lexicography*, **13**(4), 249–263, (2000).
[30] C. Lin, B. Tseng, and J. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. TRECVID Proceedings, 2003.
[31] H. Liu and P. Singh, 'Conceptnet: a practical commonsense reasoning toolkit', *BT Technology Journal*, **22**(4), 211–226, (2004).
[32] W. Maaß, 'How spatial information connects visual perception and natural language generation in dynamic environments: towards a computational model', in *Proceedings of the International Conference on Spatial Information Theory: A theoretical basis for Geographic Information Systems*, volume 988 of *Lecture Notes in Computer Science*, pp. 223–240, (1995).
[33] D. Marr, *Vision*, San Francisco: W. H. Freeman, 1982.
[34] K. McKeown, D. Jordan, B. Allen, S. Pan, and J. Shaw, 'Language generation for multimedia healthcare briefings', in *Proceedings of the Applied Natural Language Processing Conference*, pp. 277–282, (1997).
[35] G. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. Introduction to wordnet: An online lexical database. World Wide Web, (`http://www.cogsci.princeton.edu/~wn/papers.shtml`), 1993.
[36] M. Minsky, 'A framework for representing knowledge', Technical Report AI-306, Massachusetts Institute of Technology, Artificial Intelligence Lab, (1974).
[37] E. Mueller. Thoughttreasure: The hard common-sense problem and applications of common-sense. World Wide Web, (`http://web.media.mit.edu/~lieber/Teaching/Common-Sense-Course-02/ThoughtTreasure.ppt`), 2002.
[38] M. Nusseck, D. Cunningham, C. Wallraven, and H. Bulthoff, 'The contribution of different facial regions to the recognition of conversational expressions', *Journal of Vision*, **8**(1), 1–23, (2008).
[39] T. Oates, Z. Eyler-Walker, and P. Cohen, 'Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors', in *Proceedings of the International Conference on Autonomous Agents*, pp. 227–228, (2000).
[40] K. Pastra, *Vision-Language Integration: a Double-Grounding Case*, Ph.D. dissertation, University of Sheffield, 2005.
[41] K. Pastra, 'D5.1: Report on the interdisciplinary exploration of grounding levels', Public deliverable of the poeticon project (fp7-ict-215843), Institute for Language and Speech Processing, (2008).
[42] K. Pastra, 'Cosmoroe: A cross-media relations framework for multimedia dialectics', *Multimedia Systems*, (in press).
[43] K. Pastra and Y. Wilks, 'Vision-language integration in ai: a reality check', in *Proceedings of the 16th European Conference in Artificial Intelligence*, pp. 937–941, (2004).
[44] D. Roy, 'Learning visually grounded words and syntax for a scene description task', *Computer speech and language*, **16**, 353–385, (2002).
[45] D. Roy, K. Hsiao, and N. Mavridis, 'Conversational robots: Building blocks for grounding word meanings', in *Proceedings of the Human Language Technologies Workshop on Learning word meaning from non-linguistic data*, (2003).
[46] D. Roy and A. Pentland, 'Learning words from sights and sounds: A computational model', *Cognitive Science*, **26**(1), 113–146, (2000).
[47] J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. Framenet ii: Extended theory and practice. World Wide Web, (`http://www.icsi.berkeley.edu/framenet`), 2006.
[48] N. Sales, R. Evans, and I. Aleksander, 'Successful naive representation grounding', *Artificial Intelligence Review*, **10**, 83–102, (1996).
[49] S. Shapiro and H. Ismail, 'Anchoring in a grounded layered architecture with integrated reasoning', *Robotics and Autonomous Systems*, **43**, 97–108, (2003).
[50] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias, 'A visual descriptor ontology for multimedia reasoning', in *Proceedings of the workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, (2005).
[51] R. Srihari and Z. Zhang, 'Show&tell: a semi-automated image annotation system', *IEEE Multimedia*, **7**(3), 61–71, (2000).
[52] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, 'Exploiting ontologies for authomatic image annotation', in *Proceedings of the ACM Special Interest Group in Information Retrieval (SIGIR)*, pp. 552–558, (2005).
[53] L. Steels and F. Kaplan, 'Boostrapping grounded word semantics', in *Linguistic evolution through language acquisition: formal and computational models*, ed., T. Briscoe, 53–73, Cambridge University Press, (2002).
[54] R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal, 'Automating the linking of content and concept', in *Proceedings of the Association for Computing Machinery International Conference on Multimedia*, pp. 445–447, (2000).
[55] Y. Tijerino, S. Abe, T. Miyasato, and F. Kishino, 'What you say is what you see - interactive generation, manipulation and modification of 3d shapes based on verbal descriptions', *Artificial Intelligence Review*, **8**(2/3), 215–234, (1994).
[56] N. Tsagarakis, G. Metta, G. Sandini, d. Vernon, R.. Beira, J. Santos-Victor, M. Carrazzo, F. Becchi, and D. Caldwell, 'icub - the design and realization of an open humanoid platform for cognitive and neuroscience research', *International Journal of Advanced Robotics*, **21**(10), 1151–1175, (2007).
[57] S. Wachsmuth, G. Socher, H. Brandt-Pook, F. Kummert, and G. Sagerer, 'Integration of vision and speech understanding using bayesian networks', *Videre: Journal of computer vision research*, **1**, 62–83, (2000).
[58] S. Wachsmuth, S. Stevenson, and S. Dickinson, 'Towards a framework for learning structured shape models from text-annotated images', in *Proceedings of the HLT-NAACL Workshop on Learning Word Meaning from non-linguistic Data*, (2003).
[59] T. Winograd, *Understanding Natural Language*, Academic Press, 1972.
[60] S. Zinger, C. Millet, B. Mathieu, G. Grefenstette, P. H'ede, and P. Mo"ellic, 'Extracting and ontology of portrayable objects from wordnet', in *Proceedings of the MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, (2005).