Towards a Model for Grounding Semantic Composition

Marios Daoutis[1] and **Nikolaos Mavridis**[2]

Abstract.

Compositionality is widely accepted as a fundamental principle in linguistics and is also acknowledged as a key cognitive capacity. However, despite the prime importance of compositionality towards explaining the nature of meaning and concepts in cognition, and despite the need for computational models which are able to process the composition of grounded meaning there is little existing research. Thus, we aim to create computational models that concern the semantic composition of grounded meaning, that can be applied to embodied intelligent agents (such as cognitive robots), in order to make them capable of creating and processing grounded perceptual-semantic associations, and most importantly their compositions, taking into account syntactic, pragmatic as well as semantic considerations. Here we focus on an introduction to the problem, while then we review related work across multiple directions. Finally we propose a set of concrete desiderata that a computational theory of grounded semantic composition for embodied agents should satisfy, thus paving a clear avenue for the next steps towards the wider application of grounded semantics in intelligent embodied entities.

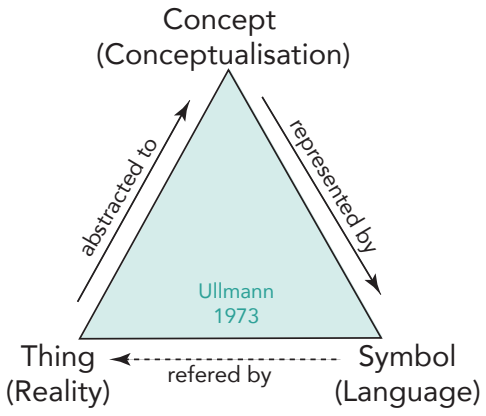

Figure 1. The Ullman Triangle – precursor of tripartite Piercian semiotics [15], and model where conceptualizations are abstractions of reality, represented in a language, explaining how linguistic symbols are related to the objects they represent.

1 Introduction

The Principle of Linguistic Relativity holds that language *influences* our cognitive processes and *affects* the way in which we conceptualize the world. It is one of the central stances connected to the intricate relation between language and thought, an idea that was originally clearly expressed by Wilhelm von Humboldt, and is usually known through the work of Edward Sapir and Benjamin Lee Whorf [10]. Given for example that a significant part of cognitive activity is marked by internal vocalization in the language we speak, which has an inherent syntax, one may ask: "what is the relation between the syntax and the semantics which in turn relate to our (often, perceptually driven) meanings?".

There exist several accounts regarding these triadic relations. In almost all of them, the principle of compositionality (also known as Frege's principle, arguably starting from Plato's Theaetetus [1]) is a central feature. According to this principle, the meaning of a *complex* expression is determined by the meanings of its *constituent* expressions and the *rules* used to combine them. This implies that concepts, such as the symbols of a language, could be structured in such a way that allows the composition of an unlimited number of complex representations (infinite meanings) from a finite number of atomic representations.

For example, linguistic compounds and modifiers are very frequent in our everyday communication even in supposedly simple expressions, e.g., "*the blue sky*" or "*the dark-red ball*". The latter example, a compound modifier (sequence of modifiers), functions as single unit where in the left-hand component, "*dark*" modifies the colour "*red*" which in turn modifies the right-hand component "*ball*". Typically modifiers can be removed without affecting the grammar of the sentence, however it is not hard to see that such omissions alter the meaning of the dependent.

Yet one perspective of meaning comes from the semiotics tradition. Bi-partite semiotics (such as ancient accounts prior to the Stoics) examine the relation between a *sign* (such as the word "*apple*") which refers to *something*, and the entity that it refers to (the *referent*). In such an account, meaning could be defined as the *relation* between the sign and the referent. Tripartite semiotics [15] add one more vertex to the picture besides the sign and referent; the concept, residing in the mind of the perceiver (see Fig. 1).

The notion of Symbol Grounding [9], which is central to

[1] Dept. of Science and Technology, Örebro University, Sweden, email: marios.daoutis@oru.se
[2] National Center for Scientific Research Demokritos, Greece, email: nmav@alum.mit.edu

modern efforts towards creating embodied artificial intelligence that can have a deep understanding of natural language, is also intricately related to the semiotics viewpoint. Harnad asks: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?" and "How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?". In contrast to Harnad's symbol grounding viewpoint, in which a symbol derives its meaning through something external to the system the symbol belongs to (referent), good old-fashioned AI (GOFAI) usually resorts to within-system (internal) relations as representations of meaning.

For example, in a semantic network such as WordNet [13], the meaning of a concept (synset, in Wordnet terms) is derived from the set of relations it has with other neighbour concepts, or even the relations it has with the totality of other concepts in the semantic network (if one adopts the stance of "*semantic holism*" [17]). In this case, and in contrast to the Symbol Grounding approach, meaning is derived from within-system concept-to-concept relations (internal); while in the case of Symbol grounding, meaning is arguably derived from the across-systems' concept-to-percept-to-referent relation.

Several contributions in the context of symbol grounding address for instance lexical semantics or semantics of referring expressions, and despite the active interest in the area, very few contributions concern grounding in relation to the compositionality of meaning. In order to generalize and radically extend the scope of symbol grounding to arbitrary expressions, an operationalisation of semantic compositionality is essential. Furthermore, in traditional linguistics, two of the main ways to approach the interpretation of meaning are: (*a*) as being dependent on the literal word meaning and usually represented in terms of formal symbolic representations (*Semantics*), and (*b*) as transcending the literal meaning and being highly dependent on the context: conversational as well as situational context (*Pragmatics*). Given the theoretical landscape above, we further extend to behavioural as well neuro-computational research, as we will seek to first review relevant literature. Then, we will propose a set of desiderata for an adequate theory of computational semantic composition that can account for grounded meaning and which can be used by embodied intelligent agents and robots.

In more concrete terms, such a theory should be capable for example to learn models of modifiers, such as *colour* modifiers, through empirical exemplars. That is, given the robot has been taught what "*red*", "*green*", and "*dark-red*" mean independently through demonstration of a set of exemplars of each the three above categories, our computational theory should be able to infer a model that generalises the meaning of "*dark*" which could in turn be used to recognize "*dark-green*" objects successfully, when they are being presented to the robot. Thus, from a set of instances of examples of `"dark<color1>"`, `"dark<color2>"`, our theory should be able to derive an empirically testable model of "*dark*" that can be used to determine the meaning of `"dark <color3>"`, even if this combination was never seen before, thus performing successful generalization, and demonstrating successful semantic composition. However, as we shall see, this approach should extend far beyond the very specific case of colour modifiers to arbitrary complex

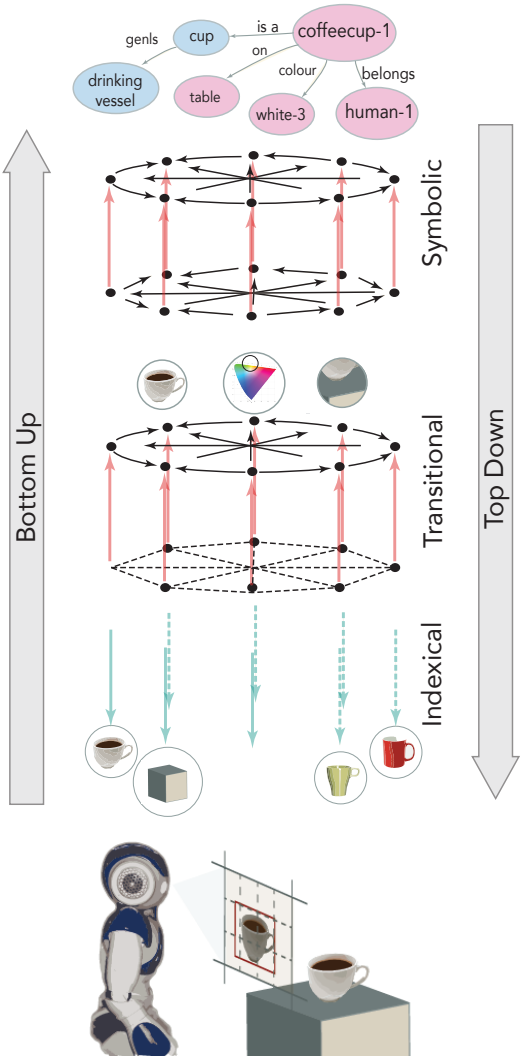

Figure 2. Illustrative scenario which depicts the fundamental forms of reference (Iconic, indexical and symbolic) when a robot is perceiving an object in its environment. Symbol relationships are composed of indexical relationships between sets of indices and indexical relationships are in turn composed of iconic relationships between sets of icons. Fig. modified from [6].

expressions, and it is only one among numerous desiderata that such a computational theory of semantic composition should fulfil. Thus, now, let us start by contextualizing our work within the background literature it belongs to.

2 Models of grounded semantic composition

Towards computational grounded composition, one of the most relevant contribution is by Gorniak and Roy [7], in the stream of work succeeded by the Grounded Situation Models proposal [12]. Gorniak and Roy present a visually-grounded model based on the combination of individual word meanings to produce meanings for complex referring expressions. Their model is based on a compositional natural language parsing framework attached to a simple synthetic vision com-

ponent, which mainly understands compositions of spatial and colour-based referring expressions. Despite the simplicity and constrained assumptions in the development of the system, the authors report success in a large percentage of test cases. However their true contribution lies into being one of the first approaches to systematically address grounded semantic composition, as well as computationally confirming that visual context affects the semantics of utterances and that the whole process of semantic composition is considerably more complex than previously thought.

In a more linguistic context, Vogt acknowledges the compositional capacity of language, as a key and distinctive aspect in understanding human cognition. While pointing to several issues related to symbol grounding and robotics, he emphasizes that when combining a holistic language (predefined structured semantics) together with learning for compositional structures, we can expect a compositional language to emerge (provided the language is transmitted through a bottleneck) [16, 19, 20]. The findings reported support his hypotheses that compositional linguistic structures can emerge when finding regularities in holistic expressions, while compositional semantic structures emerge when finding regularities in the (interaction with the) world. This is a quite interesting view on compositionality in general, because it gives to the problem a social, "language as co-evolution" context.

Another related approach is Praxicon [14], which is a resource that links natural language and sensorimotor representations of concepts, with the aim of facilitating multi-modal and multi-media content integration in cognitive systems. Furthermore, besides being heavily grounding oriented, the Praxicon advocates the compositional nature of sensorimotor representations and provides generative mechanisms for their analysis; this perspective has led to a different view of compositionality in language and the related generative mechanisms.

Grounded compositional semantics were explored also from Van Den Broeck [18], where in his paper after detailing his approach to concepts and conceptual functions, he describes a conceptualisation system intended for artificial agents, which is based on compositional meanings that emerge through networks of semantic blocks. First it is important to note the heterogeneity in the network of semantic building blocks, where perceptual and semantic data are unified. Three main, necessary components are acknowledged and presented: (a) a mechanism for grounding the concepts to the corresponding sensorimotor representations;(b) the capacity to account for semantic functions which can be thought as concept operators which take concepts as arguments and perform some operation on them in order to produce a meaning, and (c) methods to train the semantic block networks. These observations are inline with our previous work [4, 5], where we also independently acknowledged the necessity for a heterogeneous perceptual-semantic knowledge space [3]. Furthermore we elaborate on the grounding and semantic functions through our (*semantic*) grounding relations, which we use during the concept acquisition process when learning simple yet novel concepts from compositions of existing ones.

Greco & Caneva [8] study the compositionality of symbol grounding in a slightly different context, that of embodiment. In particular they studied the emerging compositional grounded representations of motor patterns, where in their empirical evaluation composed of one compositional and one holistic conditions they try to associate hand postures with words in two experiments. Surprisingly in the recognition and naming task, performance was poor for both conditions, due to the stress on the perceptual and not symbolic cues (because of the meaninglessness of the labels). However the performance of the compositional group increased considerably with the introduction of meaning in the linguistic context of the sensorimotor representation and specifically when the hand posture was relevant for differentiating between stimuli. The authors interpret this finding as emerging perceptual representations which work compositionally as a ground for the corresponding symbols, similarly to what happens with symbolic composition. This interpretation is one more evidence towards the assumption that the same cognitive mechanisms that guide the compositionality of symbols and language, to also guide the perceptual and sensorimotor compositionality. The study concludes with advocating that grounding associations between percepts and verbal concepts appeared to work both in top-down and bottom-up ways, while contributing to each others representations[1].

On the embodied end, we also see work by Chuang and collaborators [2], where novel higher level composite knowledge emerges, in the context of visual recognition and action learning. The system is based on imitation and back-prop learning which takes place when the robot is trained to learn the representations of action words, object categories and grounded natural language understanding. Still in the embodied perspective, Mangin & Oudeyer [11] stress the multi-modal aspect of grounding in learning joint word and gesture representations as well as their semantic compositions. Their experiments show that their system was able to learn the emerging semantic associations, surprisingly without providing it with a symbolic representation of the semantics.

3 Desiderata for a theory of grounded semantic composition

Following our previous discussion, we present several desirable features and attributes that we want to base our model on. With this incomplete yet approximately orthogonal set of desiderata we aim for a more general plan towards the further application of grounded semantic composition.

Perceptual-semantic-symbolic integration This is one of the most important attributes of the model as the main goal is to connect linguistic compositions and their semantic counterparts to the perception of the world. In principle this integration can be achieved through associating the multi- and cross- modal perceptual representations to the corresponding semantic representations (i.e. grounding).

Learning The underlying mechanism of the associations described above has to be mediated by empirical learning through examples.

Context, Hierarchy & Recursion These three features mainly originate from linguistics and are common in all natural languages. Therefore a model that is capable of dealing with semantic compositions should account for all

[1] As the authors mention "*symbols become meaningful on the sensorimotor grounds, but also analog representations (e.g., trajectories and postures) become more distinguishable when a specific label is available for them.*"

three characteristics since language affects the semantics of utterances.

Similarity – Dissimilarity In order to be able to form compositions, we have to implicitly have the capacity to be able to distinguish between near concepts which stand for separate entities. For example, a "*red ball*" might be almost identical to a "*dark red ball*", however we need to exploit similarity metrics in order to differentiate between the two.

Concept extension A desideratum for any acceptable computational theory of grounded semantic composition should be that every concept carries with it the processes that enable its composition with any other concept (also for not permissible), so that the result has a computable and demonstrable extension in the Fregian sense.

Polysemy & Homonymy Polysemy is when a sign (e.g., a word, phrase) has multiple related meanings (sememes). Homonymy on the other hand occurs when the multiple meanings of a word are semantically unrelated. The model should be able to handle both cases.

Conceptual alignment and drift Conceptual alignment takes place when two communicating entities need to negotiate their models of meaning in order to facilitate co-reference. Longer-term concept drift is also inevitable due to the dynamic, constantly evolving nature of the world. Therefore the properties of a concept change (sometimes continuously) over time, and their predictions are bound to become less accurate as time passes. The model should have the capacity to account for both concept drift and alignment.

4 Conclusion

In this paper we have discussed the need for a computational theory of semantic composition that can be used by embodied intelligent entities, such as robots. After motivating and contextualizing the need for such a theory, we reviewed relevant existing literature. Most importantly, we then proposed a non-exhaustive yet adequate set of desiderata that any satisfactory theory of grounded semantic composition should fulfil. Towards our ultimate dual goal of a deeper theory of embodied meaning, as well as intelligent artificial entities that can perform wide-ranging language understanding and human-robot interaction, such a theory constitutes a catalytic major step.

Acknowledgements

The research leading to these results has received partial funding from the Greek General Secretariat for Research and Technology and from the European Regional Development Fund of the European Commission under the **Operational Programme "Competitiveness and Entrepreneurship"** *OPCEII* - Priority Axis 1 **"Promotion of innovation, supported by research and technological development"** and under the Regional Operational Programme Attica - Priority Axis 3 **"Improving competitiveness, innovation and digital convergence"**.

REFERENCES

[1] L. Campbell, *The Theaetetus of Plato*, Oxford University Press, 1861.

[2] Li-Wen Chuang, Chyi-Yeu Lin, and Angelo Cangelosi, 'Learning of composite actions and visual categories via grounded linguistic instructions: Humanoid robot simulations.', in *IJCNN*, pp. 1–8. IEEE, (2012).

[3] Marios Daoutis, *Knowledge Based Perceptual Anchoring : Grounding percepts to concepts in cognitive robots*, Ph.D. dissertation, Örebro University, School of Science and Technology, Örebro University, Sweden, 2013.

[4] Marios Daoutis, Silvia Coradeschi, and Amy Loutfi, 'Towards concept anchoring for cognitive robots', *Intelligent Service Robotics*, **5**, 213–228, (2012).

[5] Marios Daoutis, Amy Loutfi, and Silvia Coradeschi, *Bridges between the Methodological and Practical Work of the Robotics and Cognitive Systems Communities – From Sensors to Concepts*, volume 21 of *Intelligent Systems Reference Library*, chapter Knowledge Representation for Anchoring Symbolic Concepts to Perceptual Data, Springer Publishing, 2012.

[6] Terrence W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain*, W. W. Norton & Company, 1997.

[7] Peter Gorniak and Deb Roy, 'Grounded semantic composition for visual scenes.', *J. Artif. Intell. Res. (JAIR)*, **21**, 429–470, (2004).

[8] Alberto Greco and Claudio Caneva, 'Compositional symbol grounding for motor patterns', *Frontiers in Neurorobotics*, **4**(111), (2010).

[9] S. Harnad, 'The symbol grounding problem', *Physica D: Nonlinear Phenomena*, **42**(1-3), 335–346, (June 1990).

[10] Harry Hoijer, *The Sapir-Whorf Hypothesis*, 92–105, Hoijer, 1954.

[11] Olivier Mangin and Pierre-yves Oudeyer, 'Learning semantic components from sub-symbolic multi-modal perception', in *Joint IEEE International Conference on Development and Learning an on Epigenetic Robotics (ICDL-EpiRob)*, (2013).

[12] Nikolaos Mavridis, *Grounded situation models for situated conversational assistants*, Ph.D. dissertation, Massachusetts Institute of Technology, 2007.

[13] George A. Miller, 'Wordnet: A lexical database for english', *Communications of the ACM*, **38**, 39–41, (1995).

[14] Katerina Pastra, 'Praxicon: the development of a grounding resource'.

[15] C.S. Peirce, C. Hartshorne, and P. Weiss, *Collected Papers of Charles Sanders Peirce*, Collected Papers of Charles Sanders Peirce, Belknap Press of Harvard University Press, 1935.

[16] K. Smith, H. Brighton, and S. Kirby, 'Complex systems in language evolution: the cultural emergence of compositional structure', *Advances in Complex Systems*, **6**(4), 537–558, (12 2003).

[17] Chris Swoyer, 'Relativism', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, winter 2010 edn., (2010).

[18] Wouter Van Den Broeck, 'Constraint-based compositional semantics', in *Proceedings of the 7th International Conference on the Evolution of Language (EVOLANG 7)*, pp. 338–345, (2008).

[19] Paul Vogt, 'Language evolution and robotics: Issues on symbol grounding', *Artificial cognition systems*, 176.

[20] Paul Vogt, 'The emergence of compositional structures in perceptually grounded language games', *Artificial intelligence*, **167**(1), 206–242, (2005).