A Formal Semantics for Iconic Spatial Gestures*

Gianluca Giorgolo

Utrecht Institute of Linguistics OTS, Universiteit Utrecht,
Janskerkhof 13a, 3512 BL Utrecht, The Netherlands

Abstract. In this paper I describe a formal semantics for iconic spatial gestures. My claim is that the meaning of iconic gestures can be captured with an appropriate mathematical theory of space and the familiar notion of intersecting modification. I support this claim with the analysis of some examples extracted from an annotated corpus of natural human-human interaction.

1 Introduction

The study of gestural behaviour in human communication has recently seen a rapid development, partially increased by the possibility of incorporating this knowledge in the design of embodied artificial agents for human-machine interfaces. However, to this date, the number of attempts to specify a formal framework for the analysis of gesture has been limited, and to our knowledge the only extensive attempt in this direction is the one by Lascarides and Stone [4]. In this paper, I address the same question of Lascarides and Stone, namely what the criteria that determine the semantic "well-formedness" of a gesture are, but we take a different approach. Rather than considering gestures a discourse-bound phenomenon, I assume that they contribute to communication at the meaning level. I will employ a montagovian perspective and show how we can account for their contribution to meaning formation in a way not dissimilar to verbal language. My proposal is complementary to the one of Lascarides and Stone, providing a more precise description of the mechanism of gesture meaning determination, which is left mainly unspecified in their account.

To keep things manageable, I restrict my attention to those gestures categorized in the literature as *iconic*. These gestures do not have a conventionalized meaning, but their interpretation is possible in conjunction with the interpretation of the accompanying verbal sentence. They *iconically* represent spatial or physical properties of the entities or events under discussion, in the sense that their formal appearance is determined by the spatial properties of the individuals/events under discussion. Another property that distinguishes these gestures from other typologies is the fact that they are completely independent of the lexical items they accompany. Their distribution is not tied to specific lexical

[*] I would like to thank Hannes Rieser for giving me access to the Bielefeld SAGA Corpus and Marco Aiello, Michael Moortgat, Yoad Winter and Joost Zwarts for many discussions about semantics, space and gesture.

items and similarly the lexical items they accompany are not dependent on the gestures, ruling out any deictic dimension of the gestures.

The semantics I propose is based on the notion of *iconic equivalence* and of *intersecting modification.* The former concept corresponds roughly to the relation holding between two spaces that are indistinguishable. My claim is that these two concepts are sufficient to explain a wide range of cases of gesture and speech interaction.

The paper is structured as follows: in Sect. 2 I will introduce first informally and then more precisely what I propose to be the meaning of iconic gestures; in Sect. 3 I will then outline a theory of space that capture most of the spatial information expressed in gestures and conclude in Sect. 4 by illustrating the semantics on the base of two examples extracted from an annotated corpus of spontaneous gestures.

2 Semantics

2.1 Informal Introduction

The meaning of purely iconic gestures can be analyzed in terms of two simple concepts: *iconic equivalence* and *intersectivity.* Iconic equivalence is the relation holding between two spaces that are indistinguishable when *observed* at a specific *resolution.* With resolution I mean a mathematical language that describe certain properties of a space and an associated notion of equivalence between spaces. The notion of equivalence determines the descriptive limits of the language, or equivalently the ability of the language of identifying differences in two spaces. An observation becomes then a description of a space in the mathematical language in question. For instance we can observe a space using Euclidean geometry and consider it iconically equivalent to another space if the two spaces are congruent up to rigid transformations. If we observe the same space using the language of topology we would consider it iconically equivalent to another space if there is an homeomorphism between the spaces.

The second component at the heart of the analysis of iconic gestures meaning is intersectivity. My claim is that iconic gestures can be analyzed as modifiers of the interpretation of the fragment of verbal language they accompany that contribute additional constraints to the interpretation. The constraints are expressed in terms of iconic equivalence between the space shaped by the gesture and the space occupied by the referents introduced by verbal language. The assumption is of course that a gesture combines only with semantically well-typed expressions, to which I will refer as *semantic constituents.*

The process of interpretation of a fragment of natural language accompanied by a gesture can then be visualized as in Fig. 1. The gesture (considered as a physical act) is interpreted as describing a spatial configuration, called an *iconic space.* This space is generated from the kinetic representation of the gesture by a procedure ϕ. The exact nature of this procedure is beyond the scope of this paper as it depends mainly on contextual and pragmatic factors. The semantic constituent (a string of words) is interpreted through a standard arbitrary interpretation function that associates with each word an element of a montagovian *frame of reference*. Additionally the words of the verbal language are given an interpretation also in a *spatial frame of reference*. This frame is an abstract representation of the physical space in which the individuals of the discourse universe exist. The two frames are connected by a family of mappings Loc that assign to the objects of the montagovian frame the space they occupy.

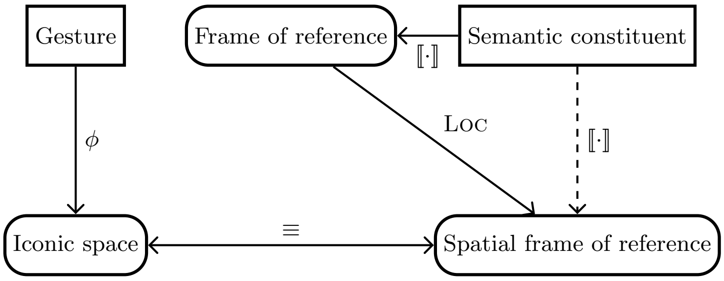


Fig. 1: Combined interpretation of speech and gesture.

2.2 Formal Semantics

As already stated, we interpret natural language expressions and gestures with respect to two types of ontologies, or frames of reference. The first frame of reference is a classical montagovian individual-based ontology $\mathcal{F}$. This frame is defined inductively as follows:

1. $D_\mathrm{e} \in \mathcal{F}$, where D_e is a primitive set of individuals,
2. $D_\mathrm{t} \in \mathcal{F}$, where $D_\mathrm{t} = \{1, 0\}$,
3. if $\varGamma \in \mathcal{F}$ and $\varDelta \in \mathcal{F}$ then $\varGamma^{\varDelta} \in \mathcal{F}$, where $\varGamma^{\varDelta}$ is the set of all functions from $\varGamma$ to $\varDelta$.

As it is the case in many semantic analyses of natural language I will assume that the domain D_e presents an internal structure that identifies sub-kinds of individuals, in particular I assume a distinction between singular and plural individuals.

The second frame of reference is a spatial ontology called $\mathcal{S}$. The frame $\mathcal{S}$ is defined inductively as follows:

1. $D_\mathrm{r} \in \mathcal{S}$, where D_r is a primitive set of regions of a space[1] equipped with some additional structure that characterizes this collection as a space (e.g.

[1] Equivalently we could use a point-based geometry. I choose here to use a region-based geometry because the logical language I propose to describe iconic spaces uses regions as primitive objects.

a relation of inclusion among regions together with the property of being an open region to consider the set a mereotopology)

2. $D_\mathrm{t} \in \mathcal{S}$,
3. if $\Gamma \in \mathcal{F}$ and $\Delta \in \mathcal{F}$ then $\Gamma^\Delta \in \mathcal{F}$.

It is important to point out that in the definition of D_r the notion of space is used in a flexible way. In most cases D_r can be considered a physical space in the classical sense but, as we will see later, sometimes we need to extend this notion to include the additional dimension of time, when for example we are interpreting gestures involving actions or events.

In what follows, we will assume the usual convention of saying that elements of D_e, D_t, and D_r have respectively type e, t and r and that elements of any domain Γ^Δ have type $\delta\gamma$.

The two frames are connected by a family Loc of (possibly partial) injective mappings from elements of $\mathcal{F}$ to $\mathcal{S}$. The elements of Loc are indexed by their domain, so for instance we will write for the member of Loc that has D_e as its domain loc_e. This implies that for each element of $\mathcal{F}$ we will allow only one mapping. We restrict the possible members of Loc with the following conditions:

1. for all $x \in D_\mathrm{e}$, $loc_\mathrm{e}(x) = r$, where r is an arbitrary element of D_p[2],
2. for all $x \in D_\mathrm{t}$, $loc_\mathrm{t}(x) = x$,
3. for all $f \in \Gamma^\Delta$, $loc_{\delta\gamma}(f) = f'$, such that $\forall x \in \Delta . f'(loc_\delta(x)) = loc_\gamma(f(x))$.

In this way the structure of the frame $\mathcal{F}$ is reflected in $\mathcal{S}$ through Loc, which is a homomorphism from $\mathcal{F}$ to $\mathcal{S}$. Also the types of $\mathcal{F}$ are reflected in the types of $\mathcal{S}$. These conditions have also the pleasant property of allowing us to define the family Loc by simply defining loc_e.

The meaning of an iconic gesture can then be expressed as a function that intersects an element of a domain in $\mathcal{F}$ with an element of the corresponding domain in $\mathcal{S}$ under the Loc mappings. We split the denotation of the gestures in two objects: a first object that inhabits a domain in $\mathcal{S}$ and that expresses the condition of iconic equivalence between the iconic space and the reference space, and a second object expressed in term of a combinator that intersects the gesture with the accompanying semantic constituent bridging in this way the interpretation of the two modes of communication.

The denotation of an iconic gestures is expressed as the characteristic function of a set of n-tuples, with $n \geq 1$, of regions such that the restriction of the space at the base of $\mathcal{S}$ to an element of this is set is iconically equivalent to the iconic space described by the gesture. Let $\rho(S, X)$ be the function that restrict the space S to its sub-region X, let $\equiv$ be the iconic equivalence relation and let γ be the iconic space associated with a gesture, we say that the denotation of a gesture g is the following function of type $r^n t$ (where with $\tau^n\sigma$ we mean a function with $n \geq 1$ abstractions of type τ):

$$[\![g]\!] = \lambda r_1 \ldots \lambda r_n . \rho \left(D_\mathrm{r}, \bigcup_{i=1}^{n} r_i \right) \equiv \gamma \ . \tag{1}$$

[2] If we choose to work with a point-based geometry then loc_e maps individuals to *sets* of points.

The combinator on the other hand acts as a glue between the interpretation of the semantic constituent and the interpretation of the gesture. We define two combinators, the first one C_{P} intersecting gestures of type $r^n t$ with constituents of type $e^n t$ (predicates) and the second one C_{M} intersecting gestures of type $r^n t$ with constituents of type $(e^n t)\, e^n t$ (predicate modifiers). The combinators also ensure that the entities depicted in the gesture co-refer with the entities introduced by natural language

$$C_{\mathrm{P}} = \lambda G.\lambda P.\lambda x_1 \dots \lambda x_n . P\, x_1 \dots x_n \wedge G\, loc_{\mathrm{e}}(x_1) \dots loc_{\mathrm{e}}(x_n) \ . \tag{2}$$

$$C_{\mathrm{M}} = \lambda G.\lambda M.\lambda P.\lambda x_1 \dots \lambda x_n . M\ P\, x_1 \dots x_n \wedge G\, loc_{\mathrm{e}}(x_1) \dots loc_{\mathrm{e}}(x_n) \ . \tag{3}$$

The application of C_{P} or C_{M} to a gesture results in an intersecting modifier in the sense of [2]. We can in fact prove the following two propositions:

Proposition 1. *Let G be the denotation of a gesture of type $r^n t$, then for every function P of type $e^n t$ we have that $C_{\mathrm{P}}\, G\, P = P \sqcap_{e^n t} C_{\mathrm{P}}\, G\, 1_{e^n t}$, where $\sqcap_{e^n t}$ is the meet operation for objects of type $e^n t$ and $1_{e^n t}$ is the unit of $\sqcap_{e^n t}$.*

Proposition 2. *Let G be the denotation of a gesture of type $r^n t$, then for every function M of type $(e^n t) e^n t$ we have that $C_{\mathrm{M}}\, G\, M = M \sqcap_{(e^n t) e^n t} C_{\mathrm{M}}\, G\, 1_{(e^n t) et}$.*

The fact that we require our combinators to correspond to the intersection (under the Loc mappings) of the meaning of the gesture and of the semantic constituent rules out the possibility of having combinators that combine iconic gestures with higher order constituent like generalized quantifiers. This restriction seems to be supported empirically by the fact that we were not capable of finding iconic gestures accompanying higher order quantifiers in a survey of a section of the Speech and Gesture Alignment (SAGA) corpus developed by the University of Bielefeld.[3]

3 A Logic for Iconic Spaces

In this short paper I will only sketch the spatial language that captures the spatial properties usually expressed with gestures. The language has been designed on the base of the analysis of the SAGA corpus. However it is probably impossible to give a general account of the spatial properties that we observe expressed in gestures, and for this reason the language has been designed to be flexible and allow the construction of different *spatial theories* for different applications. The language is inspired by various logical languages proposed in the literature, in particular the seminal analysis of Euclidean geometry by Tarski [7] and the logical interpretation of Mathematical Morphology, an image processing technique, proposed by Aiello and Ottens [1].

[3] A possible counterexample could be for example the arc-like gesture that commonly accompany a generalized quantifier like everyone or everything. However this gesture does not seem to qualify as an iconic one, given that its distribution is quite constrained to the lexical item it accompanies and moreover it is unclear which type of spatial information it is expressing.

The language is a first order language whose intended domain is the set of sub-regions of an euclidean vector space and a set of scalars. The non-logical primitives of the language are the inclusion relation ($\subseteq$) among regions, a distinguished region $\mathbf{n}$ corresponding to the points close to the origin (including the origin and) two binary operations $\oplus$ and $\odot$. The first operation $\oplus$ is defined with respect to two regions and corresponds to a generalized vector sum, known as Minkowski sum. It is defined as follows:

$$A \oplus B = \{a + b \mid a \in A, b \in B\} \ . \tag{4}$$

The second operation is defined between a scalar and a vector and is defined as follows:

$$s \odot A = \{sa \mid a \in A\} \ . \tag{5}$$

The resulting language is capable of expressing a wide range of spatial properties. It can express mereotopological properties (inclusion, partial overlap, tangential contact, etc.). The language can express the relative position of two regions (in a categorical way) by simply adding to it a number of properly defined distinguished primitive regions. It can also express relative size and with the introduction of appropriate primitives more refined comparative relations like "taller than" or "larger than". Another type of spatial feature that the language can express and that we can observe often expressed in gestures is the orientation of the main axis of a region. More in general the language is capable of expressing many size and position independent spatial properties through the use of classes of prototypes expressed as primitive regions that are scaled and translated and then used to probe the space.

To express the notion of *iconic equivalence* I will adopt a weaker version of the standard relation of *elementary equivalence* between models. I will consider two models iconically equivalent if they satisfy the same *iconic theory*. An iconic theory is simply a conjunction of atomic formulae and negations of atomic formulae. In what follows I will assume that the iconic theory has been built by the following procedure. Given a space with n distinguished regions (for instance the regions described by a gesture), we assign to each region a constant $\mathbf{r_i}$ with $1 \leq i \leq n$ and we call the set of all region constants R. Let D_{r} be the set of regions in the space, and ν the interpretation function that maps every $\mathbf{r_i}$ to the corresponding region of space, then for every k-ary predicate P we take the Cartesian product R^k and build the following conjunction:

$$\bigwedge_{t \in R^k} \begin{cases} P(t) & \text{if } S, \nu \models P(t) \\ \neg P(t) & \text{otherwise.} \end{cases} \tag{6}$$

The iconic theory is obtained by conjoining the resulting formulae.

Consequently the denotation of a gesture can be reformulated to incorporate this specific instance of iconic equivalence:

$$[\![g]\!] = \lambda r_1 \ldots r_n . \rho(D_{\mathrm{r}}, \bigcup_{i=1}^{n} r_i), \nu\,[\mathbf{r_i} \mapsto r_i] \models \Theta(\gamma) \ , \tag{7}$$

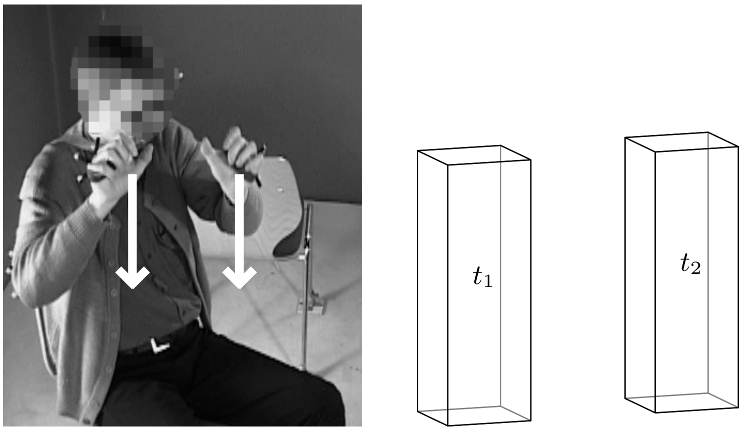
where Θ is the procedure described above for some fixed set of predicates.

4 Examples

I now analyze two examples extracted from the SAGA corpus. Beside illustrating the proposed semantics, the examples are meant to show the deep interaction between natural language semantics and gesture semantics. For this reason I selected two slightly involved cases that challenge our proposal in different ways. I will only outline the analysis of these examples: in particular I will only give an informal description of the iconic spaces associated with the gestures, as a complete formal characterization of these space would require the introduction of the complete spatial logic just sketched in Sect. 3

4.1 Interaction between Gestures and Plurals

The first example involves the interaction between plurality in natural language semantics and gestures. The example is taken from a dialogue between *Router* and *Follower*, the first describing the visible landmarks as seen during a bus ride. In the fragment we are interested in *Router* is describing a church with two towers. The speaker utters the sentence die [...] hat zwei Türme[4] ("that [...] has two towers") with an accompanying iconic gesture roughly synchronized with the noun-phrase zwei Türme. The gesture is depicted in Fig. 2 together with the associated iconic space.



(a) Gesture

(b) Iconic space

Fig. 2: Gesture accompanying the utterance die [...] hat zwei Türme and its associated iconic space

As a first step we need to define the semantics the constituent zwei Türme. To give a proper treatment of the plural Türme I assume the fairly standard

[4] The speaker is also introducing other architectonic features of the church before introducing the two towers.

extension of the montagovian frame $\mathcal{F}$ discussed in Sec. 2 consisting in the introduction of *sum individuals* (see [5]). The sum individuals are members of the type e_+ and we can know their cardinality with the function $|\cdot|$ and extract from them the individuals that compose them with a number of projection functions. I also assume a standard interpretation of a numeral like zwei as a function of type $(e_+t)e_+t$ that restrict a set of sum individuals to the subset composed by the elements with the correct cardinality (see [3]). The denotation of zwei Türme corresponds then to the set of sum individuals that have cardinality equal to 2 and that are the sum of individuals that are towers.

The proposed semantics seem inadequate to analyze this example because the number entities introduced in the verbal language does not match the number of regions depicted by the gesture (1 vs 2). However the gesture is combined in this case with a constituent referring to a plural individual and thus we can simply refine our semantics to take into account the refined individuals ontology. We extend the definition of Loc in such a way that the spatial projection of a sum individual is the tuple of the spatial projections of its composing atoms. So we say that for all $x \in D_{\mathrm{e}_+}$, $loc_{\mathrm{e}_+}(x) = \langle r_1, \ldots, r_n \rangle$, where $n = |x|$, x is the result of summing $x_1, \ldots, x_n$ and for $1 \leq i \leq n$ we have that $loc_{\mathrm{e}}(x_i) = r_i$. We also need to introduce a combinator of type $(r^n t)(e_+t)e_+t$ to intersect the interpretation of a gesture with a plural predicate:

$$C_{\mathrm{P}_+} = \lambda G.\lambda P.\lambda x.P\, x \wedge G\, \pi_1(loc_{\mathrm{e}_+}(x)) \ldots \pi_n(loc_{\mathrm{e}_+}(x)) \ . \tag{8}$$

The resulting interpretation for the noun-phrase accompanied by the gesture is the following:

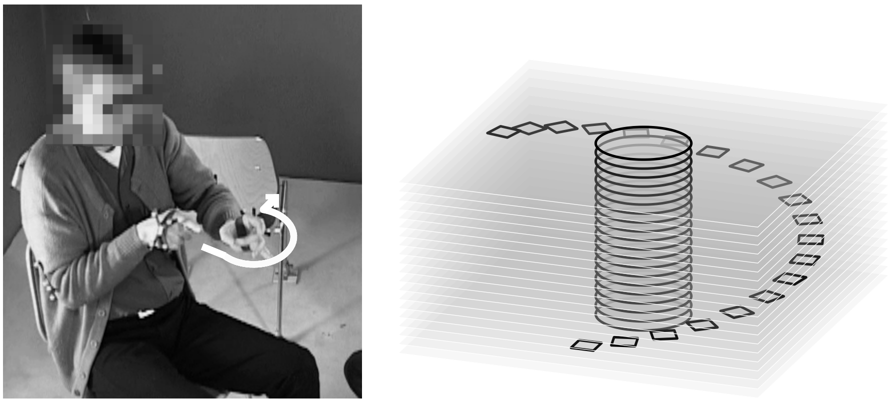
$$\lambda x.|x| = 2 \wedge \mathbf{towers}\, x \wedge \rho(D_{\mathrm{r}}, r_1 \cup r_2), \nu\,[\mathbf{r_1} \mapsto r_1, \mathbf{r_2} \mapsto r_2] \models \Theta(\gamma) \ , \tag{9}$$

where the theory $\Theta(\gamma)$ could describe for instance a space with two disconnected, vertical regions, possibly with a certain shape (e.g. a prism-like shape rather than a cylindrical one).

4.2 Gestures in the Space-Time

Quite often gestures accompany description of actions, for example by exemplifying the trajectory of a movement. The following example is aimed at showing how we can treat time in iconic gestures. My claim is that for the purposes of determining the meaning of a gesture depicting an action or an event we can consider time as an additional dimension in our spatial ontology. A realistic spatio-temporal ontology would also require additional restrictions that rule out impossible situations like objects that move with infinite velocity or that cease to exist for a certain period of time, but for the goal of demonstrating how the semantics can cope with time related issues the simple addition of time as an unrestricted dimension will suffice.

The example is taken from the same portion of the SAGA corpus. In this case *Router* explains how the bus ride goes around a pond. *Router* utters the sentence du fährst um den Teich herum ("you drive around the pond") accompanied by the gesture presented in Fig. 3. We represent the iconic space as a three dimensional space in which the vertical dimension represents time. The time dimension is "sliced" into instants to show that each instant is in itself a two dimensional space. The cylindrical region in the middle represents the constant position of the pond while the arch formed of squares represents the different positions occupied by the bus at different instants.



(a) Gesture

(b) Iconic space

Fig. 3: Gesture accompanying du fährst um den Teich herum and corresponding iconic space

The analysis of this example is in all ways similar to the analysis of the previous one. In this case I assume that the gesture combines with the predicate fährst ... herum extended by the locative preposition um.[5] The meaning of the gestures is represented as the characteristic function of a set of pairs of regions such that one represents a static circular bi-dimensional object and the other an object moving in time with and arc-like trajectory. The two regions moreover are located in the space in such a way that the circular one is roughly at the center of the trajectory followed by the other region. The set of regions satisfying these constraints is then intersected with the set of pairs of individuals corresponding to the denotation of the preposition um applied to the predicate fährst ... herum, i.e. the set of pairs of individuals such that the first one drives around the second one. In this way the referents introduced by the pronoun du and by the definite description den Teich are shared by the verb and the gesture resulting in the intuitive meaning that we would associate with this speech and gesture exchange.

[5] Nam in [6] shows how locative prepositions can be equivalently analyzed as operators that generate an intersecting predicate modifier when combined with a noun-phrase or as predicate extensors, i.e. functions that take a predicate of arity n and return as a result a predicate of arity $n + 1$.

5 Conclusion

I presented a formal semantics for iconic gestures capable of capturing what is conceivably the meaning of iconic gestures. At the moment of writing I have implemented this semantics in a speech and gesture generation prototype that can produce simple descriptions of static and dynamic space configurations that are then rendered using an animated conversational agent. I have also started testing experimentally the assumption that gesture meaning is combined with the propositional meaning of verbal language. At the same time I am also extending the semantics to treat different types of gestures in order to provide a more uniform perspective on the way verbal language is augmented by non-verbal means.

References

1. Aiello, M., Ottens, B.: The Mathematical Morpho-Logical View on Reasoning about Space. In Proceeding of the 20th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. (2007)
2. Keenan, E. L., Faltz, L. M.: Boolean Semantics for Natural Language. D. Reidel Publishing Company (1985)
3. Geurts, B.: Take Five. In Vogeleer, S., Tasmowski, L., eds.: Non-Definiteness and Plurality. John Benjamin, 311–329 (2006)
4. Lascarides, A., Stone, M.: A Formal Semantic Analysis of Gesture. Journal of Semantics (2009)
5. Link, G.: The Logical Analysis of Plural and Mass Nouns: A Lattice Theoretic Approach. In Bäuerle, R., Schwarze, C., von Stechow, A., eds.: Meaning, Use and Interpretation of Language. de Gruyer (1983)
6. Nam, S.: The Semantics of Locative PPs in English. PhD Dissertation, UCLA (1995)
7. Tarski, A.: What is Elementary Geometry?. In Henkin, L., Suppes, P., Tarski, A., eds.: The Axiomatic Method, with Special Reference to Geometry and Physics. North Holland (1959)