Applied Ontology x (200x) 1–3
IOS Press

Molecular Interactions: On the Ambiguity of Ordinary Statements in Biomedical Literature

Stefan Schulz [a,*] and Ludger Jansen [b]
[a] *Medical Informatics Research Group, University Medical Center Freiburg, Germany*
E-mail: stschulz@uni-freiburg.de
[b] *Department of Philosophy, University of Rostock, Germany*
E-mail: ludger.jansen@uni-rostock.de

Abstract. Statements about the behavior of biochemical entities (e.g., about the interaction between two proteins) abound in the literature on molecular biology and are increasingly becoming the targets of information extraction and text mining techniques. We show that an accurate analysis of the semantics of such statements reveals a number of ambiguities that have to be taken into account in the practice of biomedical ontology engineering: Such statements can not only be understood as event reporting statements, but also as ascriptions of dispositions or tendencies that may or may not refer to collectives of interacting molecules or even to collectives of interaction events.

Keywords: Bio-Ontologies, Pluralities, Dispositions, Tendencies

1. Introduction

The study of interactions between biomolecules is essential for a better understanding of biological processes, from replication and expression of genes to the morphogenesis of organisms. Statements such as "*Lmo-2* interacts with *Elf-2*" – with *Lmo-2* and *Elf-2* being proteins – occur in biomedical literature abstracts with a very high frequency and represent, in many cases, the core message of a scientific paper.

Biological text mining, i.e., the process of extracting structured knowledge from unstructured text [1], primarily targets statements, such as these, and there is a major interest in the text mining community in obtaining ontological support for their information and knowledge extraction activities. This is one of the reasons why numerous bio-ontologies have emerged, and the use of formal ontological criteria has been repeatedly advocated in order to facilitate the process of automatic processing of domain information.

Much basic ontology research in this area has already been done in the form of ontological investigations on material continuants, such as organs, cells, molecules [2, 3, 4]. However, there has been much less emphasis on a principled account of biologically relevant functions and processes, in spite of their relevance for the domain and the prominent space they occupy in biomedical ontologies and terminology systems (e.g., the Gene Ontology [5], the Unified Medical Language System [6], and SNOMED CT [7]). Especially the development of medical terminology systems has been mainly committed to traditions of semantic networks, lexical semantics, and cognitive science. Thus rather than being construed as describing real world entities by means of logical expressions, the focus has largely been set to *concepts* (i.e., representations of word meanings or mental representations) by means of conceptual relations. On this assumption, "*Lmo-2* interacts with *Elf-2*" would simply signify that there is some plausible linkage between the concepts (conceived of as mental representations) *"Interaction"*, *"Lmo-2"*, and *"Elf-2"*. As much as this approach might be adequate for structuring knowledge about the world by means of natural language or some kind of abstraction (e.g., in bibliographic thesauri, such as the Medical Subject Headings [8]),

*Corresponding author: Stefan Schulz, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany. Both authors contributed equally to this work.

it fails where exact statements and reasoning about biological entities, such as molecules, functions, or pathways are required.

Interestingly enough, scientists and other human agents are perfectly able to communicate by means of such statements, and they obviously agree on the referents (the entities in the world) which are denoted by linguistic expressions like "*Lmo-2* interacts with *Elf-2*". It seems that the semantic underdeterminations of these statements are of minor importance for their communication. From a formal ontology perspective, however, one has to take into account that "*Lmo-2* interacts with *Elf-2*" may have more than one possible interpretation and thus more than one formalization in, say, first order predicate logic. In formally representing the meaning of such statements, we have thus to make explicit what the speakers or authors precisely wanted to express.

In this paper we will demonstrate that even the formalization of an apparently simple but prototypical statement about protein interaction like "*Lmo-2* interacts with *Elf-2*" may uncover totally different ontological presuppositions.

2. Basic Ontological Assumptions

It is widely recognized that the construction of biomedical ontologies should obey strict logical and ontological criteria. To this end, top-level ontologies have been devised, such as the domain-independent ones DOLCE [9], BFO [10], and GOL [11] on the one hand, and domain-specific ontologies such as Simple Top Bio [12], GFO-Bio [13] , and BioTop [4] on the other hand. These ontologies mainly coincide in their fundamental division between continuants (endurants, e.g., material objects, spaces, qualities) and occurrents (perdurants, e.g., actions, events, processes). The distinction is that occurrents have temporal parts, which means that they are never fully present at a given time. Additionally, they are existentially dependent on continuants, i.e. they cannot exist without the participation of continuants.

Continuants are split into independent and dependent ones. Examples of *independent* continuants are material objects and spaces. In contradistinction, *dependent* continuants are entities which presuppose other entities in which they inhere. Thus they are ontologically dependent on their bearer. Examples of dependent continuants are masses, colors, and tendencies: A particular mass may inhere in a particular molecule, a particular color may inhere in a particular flower. The tendency to divide may inhere in a cell and the tendency to relieve headache may inhere in an aspirin tablet. Tendencies are related to occurrents by the relation of *realization*. They are special kinds of dependent entities, in that they need not be realized in order to exist. There are cells which never divide, and aspirin tablets that never relieve a headache.

Our ontological framework for describing molecular interaction patterns includes entities of all these kinds. For instance, a protein molecule, which is a material continuant, has a disposition to perform a certain function, e.g., binding, which is a dependent continuant. Finally, an actual realization of this disposition, *viz.* the process of interaction between protein molecules, is an occurrent.

Aware of the need to comply with existing standards for ontologies, especially in the light of the Semantic Web and the various specifications of Description Logic (DL) [14], we keep our logic simple. So we refrain from higher-order logics, as well as temporal or modal logics. We also use a parsimonious set of relations, following the OBO (Open Biological Ontologies) recommendation [15]. As a primitive formal relation we introduce the irreflexive, non-transitive and asymmetric instantiation relation **instance_of** which relates particular entities to the universals (types, kinds) they instantiate. In addition, we need a formal relation for class subsumption between universals, expressing scientific findings about relations between the kinds of things that are in the world. As scientific laws are meant to range not only over all present instances of a given kind, but also over all past and future instances and, moreover, also over merely possible instances [16], such a relation is not easy to define. We will here follow the OBO standard and introduce, to this end, the taxonomic subsumption relation *Is_a* by means of the **instance_of** relation [15]. (Throughout this paper, we use *Capitalized Initial Letters and Italics* for the names of relations between universals as well as for the names of universals. Particulars are pointed out by lower case initial letters, **bold face** is used for relations involving particulars.)

If we use variables S, S_1, ... to range over universals of independent continuants (a.k.a. Aristotelian substances); P, P_1, ... to range over universals of dependent continuants (a.k.a. properties); E, E_1, ... to range over universals of occurrents (a.k.a. events), and t to range over points of time, we can define the *Is_a* relation as follows:

$$Is\_a(S, S_1) =_{def} \forall x : (\textbf{instance\_of}(x, S) \rightarrow \textbf{instance\_of}(x, S_1)) \quad (1)$$

$$Is\_a(P, P_1) =_{def} \forall x : (\textbf{instance\_of}(x, P) \rightarrow \textbf{instance\_of}(x, P_1)) \quad (2)$$

$$Is\_a(E, E_1) =_{def} \forall x, t : (\textbf{instance\_of}(x, E, t) \rightarrow \textbf{instance\_of}(x, E_1, t)) \quad (3)$$

While *Is_a* is a subsumption relation that holds between universals belonging to the same ontological category, there are other relations that hold between pairs of universal of entities which pertain to different categories. In this paper, we will make use of a relation *Has_property* between universal independent continuants and universal dependent continuants, in this order. We define this relation with the help of the inherence relation **inheres_in** that holds between individual independent continuants and individual dependent continuants, in this order:

$$Has\_property(S, P) =_{def} \forall x : (\textbf{instance\_of}(x, S) \rightarrow \exists y : \textbf{instance\_of}(y, P) \wedge \textbf{inheres\_in}(y, x)) \quad (4)$$

Further down in the paper we will discuss a probabilistic variant of this relation.

Furthermore, we make the following ontological subdivision: When we deal with things of a certain kind, we have to distinguish between *individuals* belonging to this kind and *collectives* of individuals that belong to the same kind [2]. This very natural ontological distinction, which is mirrored by the singular / plural division in most natural languages, must be addressed wherever collectives or pluralities of individual objects occur. However, this distinction is often obscured when referring to mass entities (e.g., water vs. water molecules). Given the atomicity of material continuants, we do not admit material mass entities in our present framework, but regard them as collectives of particles instead.

A collective is given, e.g., by all *Lmo-2* molecules involved in an experiment, as opposed to exactly one individual *Lmo-2* molecule. In the following, we will use the subscript "COLL" to refer to collectives. Thus, for each universal X we principally admit the existence of a corresponding collective X_{COLL}, the extension of which is the class of collections of instances of X. For instance, *ProteinMolecule*$_{COLL}$ denotes the class of collectives of protein molecules as well as *Lmo-2*$_{COLL}$ the class of collectives of *Lmo-2* molecules. We also admit collectives of occurrents, such as *Interaction*$_{COLL}$, the class of collectives of interaction events.

3. Competing interpretations I: Event Readings

Let us come back to our example: "*Lmo-2* interacts with *Elf-2*". Such statements are generally formulated by researchers who collect scientific evidence by empirical observations. Observations of this kind are commonly made in an indirect way, since the objects under scrutiny are below the threshold of visibility. For this reason measurement procedures of varying degrees of sophistication are applied, the results of which can be used to draw conclusions about the significance of an experiment. These conclusions may vary in their degrees of certainty. This certainty is affected by measurement errors as well as by errors in the design of the experiment which then may lead to false conclusions.

Before discussing the differences between different possible readings of our sample sentence, we want to point out some elements that all interpretations share. There are, of course, several kinds of biomolecular interactions, like binding, inhibition, activation, and transport. Without subscribing to any specific theory of (bio)chemistry we can state that they all involve (1) at least two biomolecules and (2) the spatial vicinity of these, which leads to (3) a causal influence that they exert on each other. This is common ground for all these different kinds of interaction.

If a statement like "*Lmo-2* interacts with *Elf-2*" is being uttered in a laboratory or written in a scientific paper or textbook, the minimal thing that can be inferred is that there are molecules of type *Lmo-2* and *Elf-2*. As we are interested here in the scientific context only, we assume in what follows as a necessary implication of any utterance of our sample statement that there are universals *Lmo-2* and *Elf-2*. *Lmo-2* is a *Protein Molecule*, as is *Elf-2*, and *Protein Molecule* is, of course, a *Molecule*, which is an *Independent Continuant*.

Within an Aristotelian theory of universals, universals exist if and only if they are instantiated; that is, the existence of the universals *Lmo-2* and *Elf-2* implies that there are individual molecules that instantiate these universals and *vice versa*. (Whoever has a different theory of universals may have to add this as a further assumption.) All the possible readings of "*Lmo-2* interacts with *Elf-2*" to be discussed in the remainder of this paper have thus the following as a common ground:

$$\begin{aligned} &Is\_a(Lmo\text{-}2, ProteinMolecule) \wedge Is\_a(Elf\text{-}2, ProteinMolecule) \wedge \\ &Is\_a(ProteinMolecule, Molecule) \wedge Is\_a(Molecule, IndependentContinuant) \wedge \\ &\exists l, e : (\textbf{instance\_of}(l, Lmo\text{-}2) \wedge \textbf{instance\_of}(e, Elf\text{-}2)) \end{aligned} \tag{5}$$

Despite this common ground, the sentence remains highly ambiguous even within a scientific context. First we will discuss interpretations of "*Lmo-2* interacts with *Elf-2*" that interpret it as a report of an event or as a report of several events. Here are some possible interpretations of the sample statement that belong to this group:

1. One individual *Lmo-2* molecule interacts with one individual *Elf-2* molecule.
2. A collective of *Lmo-2* molecules interacts with one individual *Elf-2* molecule.
3. One individual *Lmo-2* molecule interacts with a collective of *Elf-2* molecules.
4. A collective of *Lmo-2* molecules interacts with a collective of *Elf-2* molecules.

Our sample statement may describe the fact that exactly one such interaction happened. Alternatively, it can describe the fact that a multitude of such interactions (as described in 1-4) happens, which would be the normal thing in many biochemical contexts. This adds up to eight different interpretations. But any of these interpretations is still ambiguous in a very important respect. With each of these interpretations, the speaker may mean either that such interactions did actually happen, or the speaker may mean that the molecules in question have the disposition or the tendency to interact in such a way. This gives way to even more possible interpretations. Thus, "*Lmo-2* interacts with *Elf-2*" turns out to be a highly ambiguous sentence. We will now discuss the different possible interpretations of this sentence in turn and suggest methods for representing them formally.

3.1. Occurrents involving individual continuants

On the first interpretation, "*Lmo-2* interacts with *Elf-2*" describes the fact that an individual *Lmo-2* molecule interacts with an individual *Elf-2* molecule. A standard way to render such a situation formally would be the use of the existence quantifier of first order predicate logic. For the sake of simplicity, we suppress time indices here:

$$\exists l, e : (\textbf{instance\_of}(l, Lmo\text{-}2) \wedge \textbf{instance\_of}(e, Elf\text{-}2) \wedge \textbf{interacts}(l, e)) \tag{6}$$

This formalization ensures that there is *at least* one individual *Lmo-2* molecule which interacts with *at least* one individual *Elf-2* molecule at *at least* one instant. This interpretation can now be modified in various ways. We could, e.g., add exclusivity postulates like in Formula 6 that ensure that *exactly* one individual molecule of each kind are interacting with each other. Though such a solitary event might be rarely observed in experiments, there may be contexts where this is the intended meaning:

$$\exists l, e : (\textbf{instance\_of}(l, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e, \mathit{Elf\text{-}2}) \wedge \textbf{interacts}(l, e) \wedge \forall l^*, e^* : ((\textbf{instance\_of}(l^*, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e^*, \mathit{Elf\text{-}2}) \wedge \textbf{interacts}(l^*, e^*)) \rightarrow (l^* = l \wedge e^* = e))) \quad (7)$$

Normally, however, this formalization will be too strong an interpretation of our sample statement. For any statement of this form will in fact be false, if at any other time another *Lmo-2* molecule interacts with an *Elf-2* molecule – or if at the very same time another *Lmo-2* molecule interacts with an *Elf-2* molecule at any other place. Therefore, we do not consider it as a useful interpretation of our sample sentence. We will, however, refer back to this formula and the exclusivity clauses used in it in the following discussion.

In Formulae 6 and 7 we have expressed the interaction event by means of a binary relation **interacts** between individual continuants. This relation on the level of instances is irreflexive (nothing ever interacts with itself), symmetric, and non-transitive. The OBO (Open Biological Ontologies) relation ontologies, however, recommends to restrict ourselves to a parsimonious array of basic relations. Therefore, we will eliminate the **interacts** relation, using the technique introduced by Davidson [17] to quantify over events. (To use this device for the analysis of collective plural phrases was originally suggested by [18] and has been elaborated by [19], [20], [21].) This means that we represent the interaction process as an occurrent entity in its own right rather than by the relation **interacts** as in Formulae 6 or 7. This move is made possible through our admission of occurrent entities, and it corresponds to common practice in biomedical ontologies. The relation between the particular process and the participating particular continuants is then given by the relation **has_participant** [15]. The **has_participant** relation is a relation between a particular occurrent and a particular continuant, in this order. It is irreflexive, asymmetric, and non-transitive. For nothing participates in a continuant and no occurrent participates in anything. Again, we dispense with a time index for sake of simplicity. Within a fully-fledged implementation, a time index should be included, as an occurrent may have different participants at different stages.

$$\exists l, e, i : \textbf{instance\_of}(l, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e, \mathit{Elf\text{-}2}) \wedge \textbf{instance\_of}(i, \mathit{Interaction}) \wedge \textbf{has\_participant}(i, l) \wedge \textbf{has\_participant}(i, e) \quad (8)$$

This formalization makes it easier to represent occurrents with more than two participants, as with the representation in Formula 6, where we would have to deal with n-ary relations for n participants. According to this formal representation, "*Lmo-2* interacts with *Elf-2*" is to be understood as stating that there is at least one interaction process, in which at least one protein molecule of the kinds given is involved. It does not exclude that other molecules are involved in this very interaction process. If we want to secure that *Lmo-2* and *Elf-2* are the only participants of the molecular interaction, we have to employ exclusivity conditions similar to Formula 7:

$$\exists l, e, i : (\textbf{instance\_of}(l, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e, \mathit{Elf\text{-}2}) \wedge \textbf{instance\_of}(i, \mathit{Interaction}) \wedge \textbf{has\_participant}(i, l) \wedge \textbf{has\_participant}(i, e) \wedge \forall x : (\textbf{has\_participant}(i, x) \rightarrow (\textbf{instance\_of}(x, \mathit{Lmo\text{-}2}) \vee \textbf{instance\_of}(x, \mathit{Elf\text{-}2})))) \quad (9)$$

If we want to keep the requirement of pairwise interaction, we have to add uniqueness conditions in the fashion of Formula 7 for this purpose:

$$\exists l, e, i : (\textbf{instance\_of}(l, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e, \mathit{Elf\text{-}2}) \wedge \textbf{instance\_of}(i, \mathit{Interaction}) \wedge \textbf{has\_participant}(i, l) \wedge \textbf{has\_participant}(i, e) \wedge \forall x : (\textbf{has\_participant}(i, x) \rightarrow (\textbf{instance\_of}(x, \mathit{Lmo\text{-}2}) \vee \textbf{instance\_of}(x, \mathit{Elf\text{-}2})) \wedge \forall l^*, e^* : ((\textbf{instance\_of}(l^*, \mathit{Lmo\text{-}2}) \wedge \textbf{instance\_of}(e^*, \mathit{Elf\text{-}2}) \wedge \textbf{has\_participant}(i, l^*) \wedge \textbf{has\_participant}(i, e^*)) \rightarrow (e^* = e \wedge l^* = l)))) \quad (10)$$

In contrast to Formula 7, such a formalization that quantifies over events is still much more realistic, because its truth is compatible with more than one interaction process of the same kind happening at the same time or at other times.

3.2. Occurrents involving collectives of continuants

As mentioned above, it is important to distinguish between individuals of a kind and collectives of individuals of that kind. Rector and Bittner [2] have accounted for this by introducing the formal relation **has_grain** which relates a collective c to each of its constituents e. In [22] this account has been further developed by introducing a collective universal X_{COLL} whose instances are constituted by two or more constituents which are instances of X:

$$\forall c : \textbf{instance\_of}(c, X_{COLL}) \rightarrow \exists e_1, e_2, ..., e_n, n > 1 : \bigwedge_{\nu=1}^{n} \textbf{instance\_of}(e_\nu, X) \wedge \textbf{has\_grain}(c, e_\nu) \quad (11)$$

Note that **has_grain** is a subrelation of **has_part**. As a consequence, we identify a collection as a mereological sum of its constituents (regardless of their spatiotemporal arrangement), and not as a mathematical set. The reason for rejecting the set approach is two-fold. Firstly, because mathematical sets are extensional and therefore not robust with regard to gain and loss of constituents. Secondly, because collectives should be of the same ontological category as their constituents: A collective of material objects should be a material object, and a collective of events should be an event. Sets, however, are abstract objects that do neither exist in space nor in time. We do not use **has_part** for this purpose, because participants in interactions may have parts that do not themselves participate in the interaction. A *Lmo-2* molecule, e.g., may participate in an interaction without every of its electrons being a participant in this interaction. Whereas **has_part** is transitive, **has_grain** is not. It is a irreflexive, asymmetric, and intransitive relation that holds between particular collectives and individuals.

We therefore modify our formalism substituting individuals by collectives:

$$\exists l, e, i : \textbf{instance\_of}(l, \mathit{Lmo\text{-}2}_{COLL}) \wedge \textbf{instance\_of}(e, \mathit{Elf\text{-}2}_{COLL}) \wedge \textbf{instance\_of}(i, Interaction) \wedge \textbf{has\_participant}(i, l) \wedge \textbf{has\_participant}(i, e) \quad (12)$$

It has been objected against this style of formalization that it leads into Russell type paradoxes, once we consider an event "in which all and only the members of the sets of events which do not play a part in themselves play a part" (cf. [21], p. 301). Our solution, however, is not prone to this paradoxical result, because in our ontological framework only continuants can be participants of occurrents, and thus no occurrent can be a participant of another occurrents or of itself.

3.3. Collectives of occurrents

Formulae 10 and 12 use the same type of *Interaction* occurrents for different scenarios: In the first case, a particular interaction has individual protein molecules as participants, in the second case collectives of molecules. This ambiguity is somewhat hidden through our use of a highly general term like "interaction", and it might even be irrelevant for a lot of cases. If, for example, an interaction between solutes and solvents occurs in a solution, this interaction necessarily involves collectives of both solvents and solutes as well as a collective of sub-events. In other cases both types of collectivity have to be kept apart, especially if more specific terms like "binding" are used. For a non-collective binding process can only happen between two individual molecules, not between two collectives of molecules. Thus, if we encounter a plurality of bindings within a plurality of molecules, it would not be admissible to describe this as a binding between two collectives of molecules but rather a collective of bindings between pairs of instances of the kinds of molecules in question. Thus we are confronted with a collective of processes rather than with collectives of continuants.

In order to represent such a situation, let us first introduce the collective interaction universal I_{COLL} which is constituted by individual constituents which are instances of I, analogously to Formula 11. Then

we have to determine how each of the grain interactions look like. If they are pairwise interactions between an *Lmo-2* molecule and an *Elf-2* molecule, each of these interactions fits Formula 10. Combining Formulae 10 and 11, we get:

$$\begin{aligned}\exists p, i_1, i_2, ...i_n, n > 1 : \bigwedge_{\nu=1}^{n} (&\textbf{instance\_of}(i_\nu, I) \wedge \textbf{has\_grain}(p, i_\nu) \wedge \\ &\exists l_\nu, e_\nu : \textbf{instance\_of}(l_\nu, Lmo\text{-}2) \wedge \textbf{instance\_of}(e_\nu, Elf\text{-}2) \wedge \\ &\quad \textbf{has\_participant}(i_\nu, l_\nu) \wedge \textbf{has\_participant}(i_\nu, e_\nu) \wedge \\ &\quad \forall x : (\textbf{has\_participant}(i_\nu, x) \rightarrow \textbf{instance\_of}(x, Lmo\text{-}2) \vee \textbf{instance\_of}(x, Elf\text{-}2)) \wedge \\ &\forall l^*_\nu, e^*_\nu : ((\textbf{instance\_of}(l^*_\nu, Lmo\text{-}2) \wedge \textbf{instance\_of}(e^*_\nu, Elf\text{-}2) \wedge \textbf{has\_participant}(i_\nu, l^*_\nu) \wedge \\ &\quad \textbf{has\_participant}(i_\nu, e^*_\nu)) \rightarrow (e^*_\nu = e_\nu \wedge l^*_\nu = l_\nu)))\end{aligned} \tag{13}$$

4. Competing Interpretations II: Dispositional Readings

According to the above interpretations, our sample sentence reported the existence of one or more interaction events. However, messages of the style "*Lmo-2* interacts with *Elf-2*" very often do not focus on the accidental occurrence of events but are rather meant to express some inherent property of the objects under investigation. On the one hand, it is likely that a biologist would mean "An interaction between *Lmo-2* and *Elf-2* happened" while describing the outcome of a specific experiment. On the other hand, a biology textbook would rather want to communicate something like "*Lmo-2* molecules have the disposition or tendency to interact with *Elf-2* molecules". This ambiguity, of course, matches Aristotle's famous distinction between act and potency, and Aristotle himself observed that "potency" is in itself an ambiguous term [23]. Thus the ambiguity of our sample statement increases even more, because the dispositional reading of our sample sentence is ambiguous in itself. Obviously, such a reading of "*Lmo-2* interacts with *Elf-2*" is intended to ascribe some causal or statistical property, a disposition or tendency. Thus in this case the statement should be represented as a property ascription, i.e., by means of the *Has-property* relation that we introduced for this purpose.

But even if this is the common ground of the dispositional reading, three questions remain open and have to be answered:

1. Which event is it exactly that the property in question is meant to cause?
2. What is thought to be the bearer of this property?
3. Which kind of property is in fact intended to be ascribed?

The first question can be asked because dispositions and tendencies are causal properties. They are related to certain dependent continuants or occurrences that are considered to be their realizations. Thus, the first question can be answered by pointing to the realization of the disposition in question, which will be one of the many event readings we discussed (and formalized) thus far. Our answer to the second question will at least in part depend on our response to the first question. Are all instances of a given universal bearers of the disposition in question? Or only some of the instances? Are the individual molecules the bearers of the disposition, or rather collectives of such molecules? The third question, however, leads us in to the middle of the lively debate going on regarding the ontology of dispositions [24, 25, 26]. The dispositional properties most often discussed in the literature are so-called *surefire dispositions*: dispositions to react invariably in a certain way under specific circumstances. They are one candidate for an answer to question 3. From the point of view of knowledge representation, however, there are some problems connected with surefire dispositions. First, things may react differently in different circumstances. Thus to say that *Lmo-2* molecules have the disposition to interact with *Elf-2* molecules still leaves it open under which circumstances such an interaction will occur. We could account for this by explicitly mentioning the conditions of realization for each disposition. We may, of course, not *know* all of these conditions, but this is an

epistemic problem only. A more significant problem is that there may be infinitely many causally relevant conditions that have to be taken into account, and such an infinite list would be impossible for principled reason. We could try to circumvent this problem by adding (implicitly or explicitly) quantificational phrases like "in all circumstances" or "in some circumstances". The *all*-phrase, however, will not do. For if a certain disposition would be realized under all circumstances, it will never be unrealized. Such cases may exist (e.g., the disposition of electrons to repulse each other), but normally a disposition will only be realized under certain circumstances and not be realized under others. When we use the *some*-phrase, on the other hand, many statements about dispositions for molecule interactions will become trivial, since nearly any molecule may interact with any other molecule in some peculiar way under certain (possibly very extreme) conditions. A usual way to deal with this problem is to introduce a set of standard or normal conditions [27]. In biology, this could mean that the "disposition to interact with *Elf-2* molecules" is only ascribed to *Lmo-2* if the interaction commonly occurs under biological conditions, such as physiological pH, pressure, and temperature intervals. But the problem is not solved by referring to normal conditions. For, first, the problem that infinitely many conditions cannot be described in necessarily finite lists recurs with normal conditions. And, second, biomedical knowledge may also include the behavior of molecules in non-normal or even extreme circumstances, like low or high temperatures, exposure to intensive sunlight or atomic radiations. One option at this point would be to choose a different answer to question 3. Instead of ascribing surefire dispositions we could ascribe probabilistic dispositions, i.e., dispositions to do something (under certain circumstances) with a certain probability [28]. Such causal properties are also sometimes called "tendencies" [29] or "propensities" [30]. While with surefire dispositions a certain event will happen invariably in given circumstances, the event in question will only happen with a certain probability when a tendency is ascribed. It will, of course, be crucial to know with *which* probability the event will happen. Following standard procedures in mathematical probability theory, we can represent the quantities of the probabilities in question by real numbers between 0 and 1 satisfying the Kolmogorov axioms. In biomedical experiments, observation often yields statistical results, from which a probability can be inferred. Tendencies are thus of vital importance for the representation of biomedical knowledge [29]. There can, however, be several ontological groundings for such a probability. Suppose that we observed a hundred instances of a given universal U in situations in which all conditions C necessary for the realization R of a certain disposition D were present, but that in only fifty cases R happened, i.e., in only 50 % of all cases the disposition realized itself. There are several ontological scenarios that would explain this result. Here are two of them:

- Scenario A: Every instance of U has D, which is a tendency to R with a probability of 0.5.
- Scenario B: Every second instance of U has D, which is a surefire disposition to R; the other instances of U do not have any disposition to R.

Both of these scenarios would explain the assumed observations. Which of these scenarios we choose for our account of the observation will depend on other observations and causal assumptions. If we, e.g., know that nearly always the same instances of U display R and nearly always the same instances of U do not display R, this would *prima facie* count as a reason to embrace scenario B. If, on the other hand, we know that the same instances of U sometimes do display R and sometimes do not display R, this would *prima facie* count as a reason to embrace scenario A. For such reasoning, however, we need background assumptions about the stability of the causal properties in question: how they can be stable over time, how (if at all) they can be acquired and how (if at all) they can get lost. Last but not least, scenario B can indicate that the instances of the universal U differ in certain features, which are crucial to the ability to display R. An important example of this is the observation of modified proteins produced by mutated genes as opposed to the observation of normal (wild-type) proteins.

Considering all this, there is quite a long and complex list of entities that we implicitly refer to when ascribing a disposition or tendency to a molecule:

- independent continuants (i.e., the bearer of the disposition),
- dependent continuants or occurrents (i.e., the realization),

– quantities (of probabilities), and
– states of affairs (of realization conditions).

Besides the *Has_property* relation we need formal means to relate realizable properties with their realizations, their realization conditions and the probability of their realization. If we provisionally use a (yet to be defined) quarternary relation *Has_realization_under_conditions_with_probability* relating disposition, realization, realization conditions and realization probability, we can represent scenario A by the following formalization:

$$Has\_property(U, D) \wedge \quad (14)$$
$$Has\_realization\_under\_conditions\_with\_probability(D, R, C, 0.5)$$

Scenario B, however, is more sophisticated. For there, the realization property is 1, but not every instance of U is a bearer of the disposition in question, and thus we cannot use the *Has_property* relation as defined above. We need a *Has_property* relation, that is itself subject to a probability, i.e., a ternary relation *Has_property_with_probability*, that can be defined in terms of the **instance_of** relation as follows:

$$Has\_property\_with\_probability(U, D, P) =_{def} \quad (15)$$
$$\forall x : \textbf{instance\_of}(x, U) \rightarrow probability(\exists y : \textbf{instance\_of}(y, D) \wedge \textbf{inheres\_in}(y, x)) = P$$

Given such a definition, scenario B can be represented as follows:

$$Has\_property\_with\_probability(U, D, 0.5) \wedge \quad (16)$$
$$Has\_realization\_under\_conditions\_with\_probability(D, R, C, 1)$$

As the discussion so far points out, disposition ascriptions rely on a quite complex ontology. It would extend, by far, the expressivity of description logics commonly accepted for scalable biomedical ontologies. Adequate formal tools for their representations in the above fashion have still to be rigorously defined.

5. Conclusion

We demonstrated that sentences like "A interacts with B" exhibit indeed a wide range of ambiguity. We offered several possible analyses to formally represent the different meanings of sentences of this type. Not only can they be understood to report either events or dispositions, but both of these readings admit again for a wide variety of different interpretations. Now what is the upshot of all this? How can these distinctions be brought to work for automated knowledge acquisition and information extraction tasks in life sciences?

The necessity of logic based ontologies for these purposes has been controversially discussed [31]. So what use are our own suggestions? Which one of the many possible interpretations should we choose? An obvious strategy would be to say: Which interpretation you choose depends on the intended meaning of the particular occurrence of the sentence you deal with. For text mining applications, however, that have to digest large amounts of texts in short periods of time and with as much automation as possible, this strategy would be scarcely feasible. To cope with this situation, several strategies are conceivable. One strategy would be to choose the highest common factor of all interpretations – that what is included in all. As our discussion shows, there is not much that is shared by all possible readings of such sentences.

Moreover, there are technical constraints to be met. Current ontology languages like OWL will not be able to handle the polyadic relations used in our formulae above.[1] Moreover, none of the current biomedical ontologies has implemented the distinction between individuals and collectives. They treat individual molecules on a par with collections of molecules of the same type. In addition, they do not

[1] Remember that *Has_realization_under_conditions_with_probability* is four-adic!

distinguish between individual processes and collections of processes. What can we recommend given this situation?

If an ontology O contains a class like *Lmo-2*$_O$ but does not distinguish between individual molecules and collections of molecules, this class is in fact the class union of *Lmo-2* and *Lmo-2*$_{COLL}$, and thus *Protein Molecules*$_O$ the class union of *Protein Molecules* and *Protein Molecules*$_{COLL}$, and so on. Given this situation, we are bound to interpret our sample sentence with as little commitment as possible, i.e. as reporting at least one interaction event involving *either* individual molecules *or* collections of them. Let us call this the minimal reading. According to this minimal reading, the terms "Lmo-2" and "Elf-2" in our sample sentence denote *particular* entities that instantiate either *Lmo-2* or *Lmo-2*$_{COLL}$, or, respectively, *Elf-2* or *Elf-2*$_{COLL}$. The minimal reading is robust insofar as it remains true even if many of such events happen.

At this point we can bring in another strategy, which is to set as a standard interpretation the one that is most likely the intended meaning. Which interpretation is the most likely candidate will depend heavily on the domain we have to represent. As the domain that we target in the present paper is the representation of biochemical processes, it is unlikely that we will ever be confronted with single events of interactions between individual molecules. The reasons for this are both epistemological and pragmatic: Although such individual interactions surely do happen, they cannot be detected in the laboratory, but only collectives of interactions. We get a similar picture when we move to the domain of medical care. A practitioner will never treat a patient with an individual molecule only, but always with amounts of pharmacological substances, which can be interpreted as collections of molecules, e.g. the collection of paracetamol molecules contained in a tablet. If we restrict the minimal reading to its most likely components, we end up with interpreting our sample sentence as the report of exactly one process involving collections of molecules[2].

This way, however, important distinctions get lost. Some properties, like a molecule's structure formula, or its molecular mass, are properties of the individual molecules in questions, whereas other properties like melting temperature are properly ascribed to collections of molecules only [32]. Moreover, science does not only want to describe particulars, but also wants to state general laws or natural relations between *universals*. And this is what we would expect biology text books and life science databases to do; their task is not to report isolated events, but to represent consolidated biological knowledge that has been corroborated by many different observations. (The question when to ascribe a disposition or tendency – and which one – cannot be discussed here, but cf. [28] on this).

If a law of nature is to be understood as a relation between universals [33, 34], we must take into account that any property that is ascribed to a universal is inherited by its instances. Thus if the property of interacting with *Elf-2* is ascribed to *Lmo-2*, it must be inherited by all instances of *Lmo-2*. But not all instances of *Elf-2* perpetually interact with instances of *Lmo-2*, and the same holds for *Lmo-2*$_{COLL}$ and *Elf-2*$_{COLL}$ and their instances. Thus a universal interpretation recurring to events, such as "For each instance of *Lmo-2* there is an interaction with some *Elf-2*" can easily be discarded. However, the need for universal quantifications can be satisfied by introducing dispositions, as in "Every *Lmo-2* has the disposition to interact with some *Elf-2*." Thus we have to understand this ascription on the level of universals as the ascription of a dispositional property, which then can be inherited by the instances of the universals in question [35].

As already indicated, the choice between the two readings is context-depending. Whereas the first reading might be preferred in a report on limited observations about the behavior of biological matter in an experimental setting ("*Some* Lmo-2 molecules interacted with Elf-2"), the second reading would be preferred in a text that, based on converging scientific observations, formulates laws of nature ("*All* Lmo-2 molecules have the disposition to interact with Elf-2"). Therefore, the first one would be the more robust one for a *prima facie* analysis of scientific reports. The second one would be better suited to represent accounts of consolidated biological knowledge, e.g. in review articles, textbooks, or biological databases.

Recurring to the context of ontology use for the semantic interpretation of scientific documents as the main rationale of biological text mining, as well as the need to support automatic procedures without the

[2] This prevents the interpretation as collectives of processes such as demonstrated in 3.3.

interaction with human readers, even a choice between two alternative readings could be too complex and thus unacceptable. Then we are pressed again to search for either the common core or the most likely interpretation. And again epistemological deliberations can come to help, because we normally know about dispositions only because we have observed some of their realizations. Thus the dispositional reading can corroborate the event reading, because if we know about the disposition we most likely have observed some realization events before. But again, very important distinctions get lost if we use this work-around for the representation of scientific knowledge in life science databases.

Empirical work on relevant text corpora may be helpful to verify that these are indeed the most likely (and most helpful) solutions for the understanding of biological texts and the representation of the facts represented therein. This, however, is beyond the scope of the present paper. What we have shown here is, in any case, that there is need for a more principled account of processes and dispositions in biomedical ontologies, and that a principled and logic based ontological analysis is useful even if it goes beyond the capabilities of currently existing ontology representation languages and reasoning artifacts.

Acknowledgments:
The work on this paper was supported by the EU Network of Excellence *Semantic Interoperability and Data Mining in Biomedicine* (NoE 507505), the project *Forms of Life* sponsored by the Volkswagen Foundation, and the *Wolfgang Paul Award* of the Alexander-von-Humboldt-Foundation. We are indebted to Andrew D. Spear, Olaf Wolkenhauer, and to the anonymous referees of KR-MED 2006 and Applied Ontology for valuable comments on this paper and earlier versions thereof.

Address for Correspondence:
Stefan Schulz, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany, phone: +49 761 203 6725, e-mail: stschulz@uni-freiburg.de

References

[1] Sophia Ananiadou and John McNaught. *Text Mining for Biology and Biomedicine*. Artech House, Norwood, MA, 2006.

[2] Alan Rector, Jeremy Rogers, and Thomas Bittner. Granularity scale and collectivity: When size does and doesn't matter. *Journal of Biomedical Informatics*, 39(3):333–349, 2006.

[3] Stefan Schulz and Anand Kumar. Biomedical ontologies: What part-of is and isn't. *Journal of Biomedical Informatics*, 39(3):350–361, 2006.

[4] Elena Beisswanger, Udo Hahn, Holger Stenzhorn, and Stefan Schulz. BioTop: A life science upper domain ontology. *Applied Ontology*, 2008.

[5] Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, 2001.

[6] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2008.

[7] SNOMED. *Standardized Nomenclature of Medicine. Clinical Terms*. Copenhagen: IHTSDO (International Health Standards Development Organization), 2008.

[8] MeSH. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine, 2007.

[9] Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. Sweetening ontologies with DOLCE. In Asunción Gómez-Pérez and V. Richard Benjamins, editors, *Proceedings of the 13th International Conference – EKAW 2002*, volume 2473 of *Lecture Notes in Artificial Intelligence*, pages 166–181. Berlin: Springer, 2002.

[10] Barry Smith and Pierre Grenon. The cornucopia of formal-ontological relations. *Dialectica*, 58(3):279–296, 2004.

[11] Barbara Heller and Heinrich Herre. Ontological categories in GOL. *Axiomathes*, 14(1):57–76, 2004.

[12] Alan Rector, Robert Stevens, and Jeremy Rogers. Simple bio upper ontology, 2006. [http://www.cs.man.ac.uk/ rector/ontologies/simple-top-bio/. Last accessed: May 7th, 2008].

[13] Robert Hoehndorf, Frank Loebe, R. Poli, J. Kelso, and Heinrich Herre. GFO-Bio: A biological core ontology. *Applied Ontology*, 2008.

[14] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications*. Cambridge, U.K.: Cambridge University Press, 2003.

[15] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

[16] Karl. R. Popper. *Logic of Scientific Discovery*. Hutchinson, London, 1959.

[17] Donald Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*, pages 81–95. Pittsburgh, PA: University of Pittsburgh Press, 1967.

[18] H. N. Castañeda. Comments on Davidson's "The logical form of action sentences". In N. Rescher, editor, *The Logic of Decision and Action*, pages 104–112. The University of Pittsburgh Press, Pittsburgh, 1967.

[19] J. McCawley. The role of semantics in a grammar. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart & Winston, New York, 1968.

[20] R. Bartsch. The semantics and syntax of number and numbers. In J. Kimball, editor, *Syntax and Semantics, vol. II*, pages 55–93. Seminar Press, New York, 2007.

[21] A. Oliver and O. Smiley. Strategies for a logic of plurals. *Philosophical Quarterly*, 51:289–537, 2001.

[22] Stefan Schulz, Elena Beisswanger, Udo Hahn, Joachim Wermter, Anand Kumar, and Holger Stenzhorn. From GENIA to BioTop. Towards a top-level ontology for biology. In Brandon Bennett and Christiane Fellbaum, editors, *Formal Ontology in Information Systems. Proceedings of the 4th International Conference - FOIS 2006*, pages 103–114. Baltimore, MD, November 9-11, 2006. Amsterdam etc.: IOS Press, 2006.

[23] Ludger Jansen. *Tun und Können. Ein systematische Kommentar zu Aristoteles' Theorie dispositionaler Eigenschaften im neunten Buch der Metaphysik*. Hänsel-Hohenhausen, Frankfurt am Main, 2002.

[24] Raimo Tuomela. *Dispositions*. Reidel, Dordrecht, 1978.

[25] Stephen Mumford. *Dispositions*. Oxford University Press, Oxford, 1998.

[26] Bruno Gnassounou and Max Kistler. *Dispositions and Causal Powers*. CNRS Editions, Ashgate, Aldershot, 2007.

[27] Wolfgang Spohn. Begründungen a priori – oder: ein frischer Blick auf Dispositionsprädikate. In Wolfgang Lenzen, editor, *Das weite Spektrum der analytischen Philosophie. Festschrift Franz von Kutschera*, pages 89–106. de Gruyter, Berlin/New York, 1997.

[28] Ludger Jansen. On ascribing dispositions. In Bruno Gnassounou and Max Kistler, editors, *Dispositions and Causal Powers*, pages 161–177. Ashgate, Aldershot, 2007.

[29] Ludger Jansen. Tendencies and other realizables in medical information sciences. *Monist*, 90:535–555, 2007.

[30] Karl. R. Popper. *A World of Propensities*. Thoemmes, Bristol, 1990.

[31] Sophia Ananiadou and Jun'ichi Tsujii. Thesaurus or logical ontology, which one do we need for text mining? *Language Resources and Evaluation, Springer Science and Business Media B.V.*, 39(1):77–90, 2005.

[32] Ingvar Johansson. The ontology of temperature. In *Ursus Philosophicus. Essays dedicated to Björn Haglund on his sixtieth birthday, Philosophical Communications, Web Series, No. 32*, pages 115–124. Göteborg: Department of Philosophy, 2004.

[33] David M. Armstrong. *What is a Law of Nature?* Cambridge University Press, Cambridge, 1983.

[34] Jonathan Lowe. *The Four-category Ontology: A Metaphysical Foundation for Natural Science*. Oxford University Press, Oxford, 2006.

[35] Ludger Jansen. Dispositions, laws, and categories. A critical study of e. j. Lowe's the four-category ontology. *Metaphysica*, 8:211–220, 2007.