Contents

ARTICLES

BOOK REVIEWS

Organon F 23 (3) 2016: 280-298

Two Weak Points of the Enhanced Indispensability Argument – Domain of the Argument and Definition of Indispensability

VLADIMIR DREKALOVIĆ

Department of Philosophy. Faculty of Philosophy. University of Montenegro
Danila Bojovića bb. 81400 Nikšić. Montenegro
drekalovicv@gmail.com

RECEIVED: 05-02-2016 • ACCEPTED: 27-05-2016

ABSTRACT: The contemporary Platonists in the philosophy of mathematics argue that mathematical objects exist. One of the arguments by which they support this standpoint is the so-called Enhanced Indispensability Argument (EIA). This paper aims at pointing out the difficulties inherent to the EIA. The first is contained in the vague formulation of the Argument, which is the reason why not even an approximate scope of the set objects whose existence is stated by the Argument can be established. The second problem is reflected in the vagueness of the very term indispensability, which is essential to the Argument. The paper will remind of a recent definition of the concept of indispensability of a mathematical object, reveal its deficiency and propose an improvement of this definition. Following this, we will deal with one of the consequences of the arbitrary employment of the concept of indispensability of a mathematical theory. We will propose a definition of this concept as well, in accordance with the common intuition about it. Eventually, on the basis of these two definitions, the paper will describe the relation between these two concepts, in the attempt to clarify the conceptual apparatus of the EIA.

KEYWORDS: Platonism – Enhanced Indispensability Argument – definition of indispensability – intuition.

1. Introduction

The contemporary Platonists in the philosophy of mathematics maintain that mathematical objects have an existence. However, they do not seem to be able to provide a more detailed explanation of the nature and features of that existence. To sustain their attitude, they use various arguments. One of these is the so-called Enhanced Indispensability Argument, formulated explicitly several years ago by Alan Baker, who used the following modal syllogism (cf. Baker 2009, 613):

(1) We ought rationally to believe in the existence of any entity that plays an indispensable explanatory role in our best scientific theories.
(2) Mathematical objects play an indispensable explanatory role in science.
(3) Hence, we ought rationally to believe in the existence of mathematical objects.

It could be said that Baker's formulation is an explicit consequence of the long-term discussion held on the relation Nominalism-Platonism[1] regarding the necessity to specify the *sort* of indispensability which mathematics could treat as a scientific subject.[2] The idea behind the Argument is quite natural. Broadly speaking, if science describes and *explains* phenomena and objects which doubtlessly exist, then such a feature – an existence – must also be attributed to the tools used in those explanations. Since, among other reasons, we use mathematical objects to explain empirical phenomena, we can conclude that those objects do exist. Historically speaking, the Enhanced Indispensability Argument (henceforth EIA) is an "improved" version of the so-called Quine-Putnam indispensability argument (IA), according to which the role that mathematical objects have in describing and explaining empirical phenomena is reduced to quantification and indexing of the physical objects.[3] In addition

---

[1] See, for example, Melia (2002) for the Nominalist, and Colyvan (2002) for the Platonist side.

[2] By the word 'science' in this text, we will imply empirical sciences, such as physics, chemistry, biology, etc.

[3] The classic position in the reference books is occupied by Putnam (1971, 65). See, for example, Melia (2000, 455), Yablo (2000, 197), Colyvan (2001, 10). Nevertheless, there are authors who have been trying to prove that neither Putnam, nor Quine can be

to this, the EIA places an emphasis on the indispensability of the explanatory role of the mathematical objects in the empirical science.

The aim of this paper is to draw attention to the difficulties entailed in the EIA. The first difficulty is reflected in the vague formulation of the EIA. This vagueness is the reason why it is not possible to determine the scope of the set of objects whose existence is stated by the Argument.[4] The lack of precision, as this paper will show, even though prevalently technical in nature, reminds of that precarious and vital question which has remained unanswered since the beginnings of Platonism.[5] The other difficulty is reflected in yet another imprecision. Namely, it refers to the notion of the *indispensable explanatory role*,[6] the meaning of which had not been specified until recently, which could have resulted in different interpretations of the concept and, consequently, in different interpretations of the EIA. For this reason, the major attention will be given to the concept of (in)dispensability. More precisely, a recent proposal for the definition of the *indispensability of a mathematical object* will be recalled here; its drawback will be pointed out and a possible improvement of this definition will be suggested. Following this, the paper will deal with an unpleasant consequence of the arbitrary use of the concept *indispensability of a mathematical theory*. It is evident in the intuitively hardly graspable relationship between indispensability of an object and indispensability of a theory. We will, therefore, propose a definition of the indispensability of a mathematical theory trying to follow the line of the intuition generally held about this notion. Finally, on the basis of the two definitions – one improved and the other only suggested – we will describe the relation between these two concepts thus attempting to clarify, at least to some extent, the conceptual apparatus used in the EIA.

---

accredited with the main part of the indispensability argument. For more information, see Liggins (2008).

[4] There are opinions that scope does not matter in the case of the IA, but that what matters is a question of its specificity (cf. Baker 2003, 52). It rather seems that in the case of the EIA, as a more explicit and more precise argument, the question of scope cannot be declared as a peripheral one.

[5] Is it possible to speak of the existence of *only some* mathematical objects?

[6] For the reasons of brevity and clarity, in most cases henceforth the simple term *indispensability* of a mathematical object or mathematical theory will be used instead of *indispensable explanatory role* of a mathematical object or mathematical theory.

2. Baker's example – dilemmas

When it comes to the explanatory role of mathematics in science, there are several analyses directly related to the EIA. Some of the authors point to the impossibility of reaching any conclusion about the existence of mathematical objects on the basis of their explanatory role in science (see Bangu 2008), whereas the others claim that mathematical objects possess no explanatory capacity whatsoever in the case of empirical events (cf. Daly and Langford 2009). Also, there are authors who adhere to the standpoint that mathematical objects and models do not explain empirical phenomena in a genuine way, but only represent them in one of the possible ways (cf. Saatsi 2011), while others observe that the expression "indispensable explanatory role" has been used imprecisely in the EIA (see Molini 2014). The latter observation will be the main focus of this discussion. Let us be reminded of the famous cicada example, the common point, used to illustrate the mathematical explanation of an empirical phenomenon:

> The example featured the life cycle of the periodical cicada, an insect whose two North American subspecies spend 13 years and 17 years, respectively, underground in larval form before emerging briefly as adults. One question raised by biologists is: why are these life cycles prime? It turns out that a couple of explanations have been given that rely on certain number theoretic results to show that prime cycles minimize overlap with other periodical organisms. Avoiding overlap is beneficial whether the other organisms are predators, or whether they are different subspecies… (Baker 2009, 614)

> For example, a prey with a 12-year cycle will meet – every time it appears – properly synchronized predators appearing every 1, 2, 3, 4, 6 or 12 years, whereas a mutant with a 13-year period has the advantage of being subject to fewer predators. (Goles et al. 2001, 33)

This seems to be an example of a purely physical phenomenon being explained by mathematical tools. It applies one of the basic facts of the number theory. Since the prime number can only be divided by itself and by 1, the cicada whose life span equals a prime number has more chance of survival than the cicada whose life span equals a composite number, because the latter encounters a larger number of predators during life cycles than the former. However, it is not clear that the above example is the case where mathematical

objects play an *indispensable* explanatory role. Baker does not find it necessary to define the notion of indispensability, presumably considering its meaning as intuitively sufficiently clear. As we can see, the domain of the attribute *indispensable* is considerably broad. Mathematics can play an indispensable explanatory role in science, and so does a mathematical apparatus or a mathematical object (see Baker 2009, 613-614). The indispensability of the mathematical object *O* for the explanation of the physical phenomenon *P* is non-formally understood as the impossibility to explain the phenomenon *P* without the use of the object *O* and its accompanying features. Therefore, in this case, to explain the phenomenon *P*, no other mathematical object can be helpful. Moreover, no object whatsoever can be used for the purposes of explanation.

Before turning to the analysis of the concept of indispensability, let us accept it intuitively, as Baker did, and let us return briefly to the EIA. In the formulation of the Argument many imprecisions can be found, which could create additional confusion. What, exactly, is it about? The conclusion of the EIA tells us that we ought to rationally believe in the existence of mathematical objects. It is a rather vague formulation of a potentially very important proposition, which can create various interpretations of the EIA. Namely, it is not clear whether we ought to believe or not in the existence of *all* or just *some* of the mathematical objects. This question may seem not so important at first; however, the answer to it fundamentally determines not only the further stages in clarification of the indispensability concept and defense of the EIA, but also the consistency of the Platonist attitude on the existence of mathematical objects (see Baker 2003, 53). To answer this question, it is necessary to solve the corresponding detail in the second premise of the EIA first. In other words, we should establish whether *all* or just *some* of the mathematical objects play an indispensable explanatory role in science. It is as if Baker, as well as those who used the formulation for the purpose of analyses and criticism (see, e.g. Molinini 2014), has not omitted the potential quantifier by coincidence, leaving thus a room for the possibility of various interpretations on the one hand, perhaps for the improvements as well, and rendering all criticism easier on the other. The imprecision in the definition of the EIA, however, with its lack of quantifier, cannot be a support to Platonism.[7]

[7] If we were to express nominalist point of view with an opinion that there do not exist any abstract (mathematical) objects (see Baker 2003, 49), then the question about the EIA domain could easily be circumvented with the following answer: the primary

We cannot know with certainty what idea Baker had in mind when he formulated the EIA. Nevertheless, analyzing his famous cicada example, it may be deduced that he gravitates more towards the particular quantifier premise:

(2a) *Some* mathematical objects play an indispensable explanatory role in science;

And, consequently, towards the conclusion:

(3a) Hence, we ought rationally to believe in the existence of *some* mathematical objects.

Namely, by means of the cicada example, he illustrates the proposition about the existence of mathematical explanation in science, as well as the indispensability of mathematical apparatus, which suffices as the proof of a proposition such as *some As are Bs*. One single example, without any attempt to systematically find a role of every mathematical object used in the explanation of physical phenomena is, needless to say, far from endeavors to explain that all mathematical objects are indispensable for explaining physical phenomena. If such is the case, then we could speak about the existence of a mathematical object, more precisely, those mathematical objects that are indispensable for the explanation of physical phenomena. On the other hand, we would allow that other mathematical objects, about whose existence we do know, do not exist, and, also, that some of them do exist without our knowledge of them at the moment, since we perhaps do not know yet about a mathematical explanation of the physical phenomenon in which those mathematical objects are used. If we are to follow this line of argument, let us consider a mode to use the EIA in the cicada example. We could, for example, claim that numbers 13 and 17 exist or, to extend it, that all prime numbers exist, even though we could have given the explanation in this case without using the concept of the prime number, but using only the feature of divisibility common to all prime numbers, 13 and 17 included. If we interpret the EIA in a more flexible way, we could claim

purpose of the EIA is to refute nominalism. The existence of just one abstract mathematical object is enough to do this, hence there is a sense in which the EIA can succeed without addressing the scope. However, could we possibly accept that the main mission of the EIA is rebuttal of nominalism, without an attempt to create a systematic tool supportive of Platonism? We do not encounter a support for such a viewpoint in Baker's recent texts (cf. Baker 2005; 2009; 2015).

that there are composite numbers as well, because without comparing them with 13 and 17, we would not be able to understand the "advantages" of the prime numbers in this particular example in the first place. Nonetheless, no matter how flexibly we understand the application of the EIA, on the basis of this physical phenomenon and the EIA, we will not be able to claim the existence of some other objects of the Number Theory for certain, such as, for example, Euler's function,[8] and, in particular, those objects which do not belong do the Number Theory, such as Polish space,[9] an object of the general topology. Indeed, as the mentioned objects are not used in the specific example, and as it has not been clearly indicated that they would ever be used for explaining a physical phenomenon, we cannot speak of their existence on the basis of the EIA. We can, therefore, speak of two levels of mathematical objects: of the "privileged" ones, which exist, and of those which do not have such a status, at least not at present. Evidently, the idea to use the EIA in order to prove the existence of *just some* of the mathematical objects appears rather unsustainable and easily discardable. Similar to this, the "partial" Platonism, seen recently in the philosophy of mathematics, was short winged as well.[10]

Let us return to the concept of indispensability. Baker regarded it as intuitively clear, although he must have been well aware that the majority of objections to the EIA were to be expected on that very point. Namely, from the mathematical as well as layman's standpoint, the question highly expected is: in which way do we choose the mathematical apparatus for explaining a physical phenomenon? Is this choice an unambiguous process and what directs it? Is the whole process of selection an arbitrary one, carried out within random

---

[8] Euler's function φ maps an arbitrary natural number m into the number of integers from 0 to $m - 1$ that are relatively prime to m. For example, $\varphi(1) = \varphi(2) = 1$, $\varphi(3) = \varphi(4) = 2$, $\varphi(5) = 4$. See Erdos and Suranyi (2003, 58).

[9] A Polish space is a separable and completely metrizable topological space. There are two fundamental examples of Polish spaces. The first one is the Baire space N^N consisting of all sequences of natural numbers. The second one is the Cantor space 2^N consisting of all sequences of 0's and 1's. See Dodos (2010).

[10] In Maddy (1990) we find an extremely odd idea about existence of only those mathematical objects which have a practical application, whereas other objects' existence is denied. The author abandoned that position afterwards, as it can be seen in Maddy (1997).

circumstances, such as the affinity of the researcher, the current state of development of one of the mathematical theories, practical interests, etc.? Baker stated three types of arbitrariness that may occur in the explanation of a physical phenomenon (object, concept and theory arbitrariness), showing that none of them affects the EIA in any important way.[11] However, in addition to these three, we can point to another type of arbitrariness which, generally speaking, has often been present in the mathematical community. We will name it *isomorphic* arbitrariness. In effect, it is a mode of mathematical thinking which is expected and natural, a type of attitude for which every mathematician is prepared even during the undergraduate university education. When we analyze the content of a mathematical theory M_1, it is mathematically natural to wonder whether, perhaps, there exists another theory M_2 that would be isomorphic to the theory M_1.[12] If that is the case, then, theoretically speaking, every object, proposition, proof or explanation within the theory M_1 has its analogon in the theory M_2. It further implies that if a physical phenomenon P is explained by means of the object O_1 of the theory M_1, then it can be explained, with equal adequacy, by the corresponding object O_2 from the theory M_2. The choice of the alternative theory/object in this case does not depend on the physical phenomenon, but exclusively on the affinity of the researcher, or on some practical circumstances.[13] Which one of the objects, O_1 or O_2, is indispensable to the phenomenon P? None, according to Baker's intuition. Nevertheless, if we assume that there are no other objects which explain P, phenomenon P cannot be explained without at least one of these two objects. Therefore, they possess

---

[11] For further information, see Baker (2009, 615-619).

[12] In other words, we will say that two theories (structures) M_1 and M_2 are isomorphic if there is a bijection between them that "preserves" all the relations and operations from the domain onto the codomain. If we would want to define isomorphic vector spaces within linear algebra, then we could do it in the following way:

An isomorphism between two vector spaces V and W is a map $f: V \to W$ that

1. is a correspondence: f is one-to-one and onto;
2. preserves structure: if $a, b \in V$ then $f(a + b) = f(a) + f(b)$, and if $a \in V$ and $k \in V$ then $f(k\,a) = k\,f(a)$.

[13] Molinini pointed out the role of pragmatic circumstances in the decision-making process when it comes to choosing a suitable mathematical theory for explanation of a physical phenomenon (cf. Molinini 2014). However, this text offers the examples from alternative theory (set theory) and Minkowski geometry, which are not isomorphic in the strictly formal sense.

a kind of *common* indispensability to *P*. A partial confusion created by this example proves a need for a more precise definition of indispensability.

On the other hand, the procedure of finding mathematical explanation of a physical phenomenon is methodologically similar to the procedure of finding mathematical explanation/proof of a mathematical phenomenon/proposition. In other words, extrinsic use of mathematical tools is methodologically similar to their use in intrinsic circumstances. When we deal with a proposition that should be proved in mathematics or, more realistically, when we have an intuitive sense of the correctness of a proposition, then we start from the already proved propositions and move towards the aimed proposition. There can be many proofs of this type and we could hardly ever state that we have reached their definite number.[14] Correspondingly, in the extrinsic conditions such as Baker's cicada example we do not have formal tools by which we could prove the indispensability of mathematical objects. Namely, how can we prove that there is no other explanation within the number theory or some other theory? To put it differently, in order to state a proposition on the indispensability of the prime numbers in the cicada example, we should have a proof of the impossibility of a different mathematical explanation, which is far from a trivial task. Generally speaking, if we know that at time t_1, O_1 is the only mathematical object (also the object of the theory M_1) used in the explanation of the phenomenon *P*, we cannot state the *absolute* indispensability of the object O_1 to the phenomenon *P*. In order to state such a proposition, we ought to prove that the phenomenon *P* cannot be explained at any other time t_n, $t_n > t_1$, of the development of mathematics, by no other object O_n (which would be an object of the theory M_n). Since at moment t_1 we cannot know the explanatory capacities of objects and theories which are to be created in future, we cannot hope for such a proof either. What makes sense, however, is a consideration of a *conditional* indispensability in this context, that is, an indispensability that would aim at establishing itself as such in relation to the objects of the mathematical theories defined prior to the moment of the consideration of indispensability. Does this make the situation simpler? It does, so far as it provides a clear domain of defined objects on which indispensability is to be examined. However, broadly speaking, is there a methodology by which we could

[14] For example, several proofs of the Fermat's little theorem are known today. A more extreme example provides several hundreds of proofs of the Pythagorean theorem. See Alkauskas (2009) and Loomis (1972).

precisely solve the question of the indispensability of the object that explains a phenomenon? Is there an algorithmic set of stages that would reveal with certainty that, for instance, there is no other mathematical object, taking into account all those defined so far, in all the theories, by which we could explain the cicada example? We are not in the possession of such a methodology at present, and the indispensability which we may attribute to an object is in this sense additionally conditional and relativized. The most we can say about an object is that it is indispensable to a phenomenon unless proved differently, which is a rather discouraging position from a researcher's viewpoint. Given this situation, it is far more pragmatic and reasonable to turn to more modest aims. One of these would be: a more precise definition of the concept of indispensability.

3. Is Molinini's definition suitable?

The first part of this paper has underlined, among other things, the importance of a more precise definition of the indispensability concept within the EIA, with the aim of re-examining the power of the Argument as one of the main supporting tenants of Platonism. Daniele Molinini was the one to make a decisive and welcome attempt at this, proposing a definition of indispensability. In effect, he offered an explicit definition of *dispensability* (henceforth 'D1'):

> A mathematical entity x is explanatorily dispensable to a scientific theory T if it is possible to find a theory T^* that:
> *(a)* does not employ the vocabulary of the mathematical theory M in which x is defined;
> *(b)* offers the same (or even more) explanatory power as T;
> *(c)* is empirically equivalent to T. (Molinini 2014)

We can notice that, when compared to Baker, the domain of the predicate *is (in)dispensable to* is more precise, at least in this definition. A mathematical object x is dispensable, or not, relative to a scientific theory T. In the first position of the predicate an individual mathematical object is implied, whereas the second position is occupied by an individual scientific theory. The intuition behind this definition is clear enough and it is similar to Baker's. Informally speaking, according to the definition, the mathematical object O is explanatorily dispensable

to the theory *T* if it is possible to explain any phenomenon described by the theory *T* without *O*. Molinini provided several examples of the explanatorily dispensable mathematical objects, such as orthogonal matrices, Minkowski metric, set theory objects, etc. (see Molinini 2014), thus shattering the last hope that the EIA can be used to prove the proposition about the existence of all mathematical objects.[15]

D1 was expected to be the operative tool by means of which we could establish with certainty whether a specific mathematical object is indispensable to a specific scientific theory. Let us see if D1 reached this goal – a formalization of the concept which had been used non-formally beforehand, that is, if this formalization covered all the cases which we non-formally consider as dispensable. If we are to pursue Baker's intuition, we can state that the mathematical object *O* is explanatorily indispensable to the physical phenomenon *P* if and only if the phenomenon *P* cannot be explained without using the object *O* and its features. According to this, the mathematical object *O* is not explanatorily indispensable, that is, it is dispensable to the phenomenon *P* if and only if the phenomenon *P* can be explained without using the object *O* and its features. Therefore, intuitively speaking, a mathematical object is dispensable not only if there is an alternative to it when it comes to explaining a phenomenon, but also, as is trivially implied, if it does not explain a phenomenon at all. For example, Minkowski metric, as an object of the Minkowski geometry, is dispensable to a phenomenon of the theory of special relativity, known as FitzGerald-

---

[15] The problem of the so-called *weaker alternatives* in the explanation of phenomena is emphasized in Pincock (2012, 212-213). Claim *p* and claim *q* explain (individually) phenomenon *P*, with *p* being a stronger mathemathical claim than *q* (*q* follows from *p*, but not vice versa). If the explanatory power of claim *q*, when connected with the phenomenon *P*, is not lesser that that of *p*, it is not clear on what basis *p* would be preferred over *q*. For example, in the cicada case, let us assume that

p: prime periods minimize intersection (as compared to nonprime periods);
q: prime periods of less than 100 years minimize intersection.

According to Pincock we would be able to use *q* as an equally powerful explanation of the chosen phenomenom. In the context of the EIA, however, this quarantees existence of only those numbers smaller than 100, which is obviously an unacceptable consequence.

I am grateful to the anonymous referee who has brought my attention to this point.

Lorentz contraction.[16] We speak of dispensability in this case because in addition to the explanation of this phenomenon in which Minkowski metric is used, there is also an alternative – an axiomatization of the set theory by means of which the description of the contraction is acquired as a theorem.[17] Also, as for the Polish space, a general topology object, the same is true. Namely, on the basis of the available reference books, the Polish space is entirely unusable for explaining length contraction, which makes it dispensable to this phenomenon. As far as examples like these are concerned, D1 follows intuition. According to it, some of the objects of set theories, as well as Minkowski geometry, are not the only ones dispensable to the length contraction, but the same goes for the Polish space, being an object which does not explain it at all.

What we intend to suggest is that D1 has not covered all the intuitively dispensable objects. It is, therefore, too narrow. The problem here does not lie in the objects which explain a certain phenomenon, but for which there is an alternative explanatory object, neither in the objects which do not explain it but are part of the mathematical theory to which the object that explains the phenomenon does not belong. In these cases, D1 functions correctly. In other words, according to it, these objects are dispensable. The target of our attempts to show that this definition is not broad enough includes those mathematical objects that are dispensable on the basis of the criterion: "[it] does not explain a scientific phenomenon and belongs to the same mathematical theory as the object which does explain the phenomenon." Indeed, let us assume that x and y are objects of a mathematical theory M, the object x being enough to explain the phenomenon P of the scientific theory T, with no alternative of another object from another theory that could explain P, the object y included. Let us also assume that the object y does not explain any other phenomenon of the theory T. Intuitively, y is explanatory dispensable to the theory T since it is not used in any way to explain any of its phenomena. However, it is not dispensable according to D1, it is indispensable! How? In relation to y and T, the condition a) of the definition D1 was not fulfilled, because it is not possible to find a theory T^* which does not employ the vocabulary of mathematical theory M

[16] It is a phenomenon in which the length of the body in motion is shortened, according to the precisely set formula, depending on the velocity of its motion in relation to the point of the observer.

[17] On the proof of dispensabilty of the Metric and the use of the set theory in this case see Molinini (2014) and Andreka et al. (2007, 29-30).

in which *x* and *y* are defined and which fulfills the remaining two conditions of the definition. We can say that on the basis of D1, *y* is indispensable without taking merits for it, which is neither expected nor desirable. For example, if the numbers 13 and 17, or prime numbers in general, are indispensable objects of the number theory in the cicada example, then an object of number theory such as the previously mentioned Euler's function is altogether unusable for an explanation of the phenomenon and cannot therefore be intuitively indispensable.[18] Contrary to intuition, however, D1 gives it precisely that kind of status. Thus, D1 formally allows for a large class of objects to be considered indispensable, even though they are not intuitively experienced as such, which subverts the very purpose of defining.

Along the lines of these objections, we can propose a possible improvement of D1. It would suffice to alter only the initial part of the definition D1:

> *A mathematical entity x is explanatorily dispensable to a scientific theory T iff either x does not explain any phenomenon described by the theory T, or it is possible to find a theory T * that:*
> 1. *does not employ the vocabulary of the mathematical theory M in which x is defined;*
> 2. *offers the same (or even more) explanatory power as T;*
> 3. *is empirically equivalent to T.*[19] (henceforth D2)

In addition, D2 includes all the types of the previously mentioned cases which we understood as dispensable and D1 did not treat them as such.

Despite the fact that the difference between D1 and D2 may appear as only technical and insignificant, it turns out that it changes the conception of the indispensability of the entire mathematical theory. Before expanding on this, let us refer to a notational remark. Hereinafter, due to reasons of brevity and

---

[18] We assume here that the object in question is not used to explain another phenomenon which, along with cicada example, could be placed into a wider biological theory, such as, for instance, the cicadas' life-cycle theory or theory of the life-cycles of animals in general.

[19] In D2 we have not specifically differentiated between unexplanatory mathematical objects of the scientific theory *T*, depending on the fact if they do or do not belong to the mathematical theory whose object (possibly) explains a phenomenon of the theory *T*. A definition which would insist on such a sensibilty would probably be rather more complex and far more different from D1.

clarity, we shall refer to the mathematical object *x* which does not play an explanatory role in theory *T* at all as *trivially* dispensable. If an object *x* plays an explanatory role in theory *T*, but there is an alternative to it, another mathematical object *y*, then we shall say that *x* is *non-trivially* dispensable for *T*.

Another interesting novelty about D1 is that it considers the explanatory indispensability of a mathematical object to a scientific theory (or to a phenomenon of that theory) within the framework of a suitable mathematical theory, within which the object is defined. This approach seems correct, as the objects are defined by means of the vocabulary of the theory to which they belong. Also, the features of those objects are formulated in relation to other objects of the theory. Nevertheless, after the object-theory context in D1 was established, there is one thing which remained vague. Even though Molinini reserved the first position in the domain of *is dispensable* predicate for mathematical objects, by which he does not entail mathematical theories, he still speaks about dispensability of mathematical theory as well, asserting soon after that

> In fact, it says that dispensability of an entity is tantamount to the dispensability of the theory in which that entity is defined, and vice versa… (Molinini 2014)

This does not define dispensability of the theory at all. In this respect, Baker and Molinini take a similar position. The former employed the concept of an object's (in)dispensability in a non-formal manner, whereas the latter employed a theory's dispensability in such a way. If we attempt at questioning the justifiability of the above quotation, it ensues that we cannot treat the dispensability of a theory in a non-formal way either.

Every mathematical theory defines some mathematical objects.[20] At first, it may appear natural to state that a mathematical theory is dispensable to

[20] An additional explanation should be provided at this point, which could have been done earlier, when the notion of *isomorphic arbitrariness* was introduced. I want to thank an anonymous referee for having brought my attention to this point. Namely, when we say that *every mathematical theory defines some mathematical objects* we are in effect referring to the theory-object relation which is common in mathematics. Theory *is composed* of objects, of their features and relations that exist among them. Objects are described by means of definitions and by means of propositions. When we say *objects* we refer to all basic and defined concepts that are part of a theory. For example, prime and composite numbers, as well as Euler's function are defined objects of number

a scientific theory *T* if all its objects are dispensable, that is, to state that a theory is indispensable if it has at least one object which is indispensable to the theory *T*. However, if we are to proceed in that way, then we would, for example, consider as dispensable the mathematical theory *M* which contains the objects *x* and *y*, both being non-trivially dispensable to the theory *T*, if, in that case, there are no objects outside the theory *M* which play an explanatory role in the theory *T*. It would mean that some phenomenon described by the theory T cannot be explained without the theory *M*, and, consequently, it would not be in accordance with the intuition that instructs us to state that the theory *M* is dispensable. For that reason, we need a new definition that would follow the usual intuition about the dispensability concept, also respecting the last particular case:

> *A mathematical theory M is dispensable to a scientific theory T if and only if for every object x of the theory one of the following conditions is fulfilled:*
> 1. *The object x is trivially dispensable to the theory T;*
> 2. *The object x is non-trivially dispensable to the theory T and there is a mathematical object y which does not belong to the theory M, and which is non-trivially dispensable to the theory T* (henceforth D3).

This definition makes it clear that a mathematical object can fulfill only one of the set conditions. On the basis of this, we shall consider as dispensable only that theory in which all the elements are dispensable, with the exception that, if it is a non-trivial dispensability, we can find an alternative mathematical

---

theory. Vectors and vector spaces are objects of linear algebra. Namely, both vector and vector space belong to the category of defined objects. Neuter element of the structure (N, +) is an object of the algebra, but that structure is itself also an object of the algebra. Indeed, both the neuter element and the structure (N, +) are also defined concepts. Thus, the world of mathematical objects is rather broad and composed of various entities, not unlike the biological world of which we all are parts. This complexity does not appear to be a reason for concern because it does not entail neither formal nor intuitive obstacles related to the analyses of the EIA. Let us mention that every mathematical object is observed in *the context* of some theory or, more specifically, in the context of some structure. For instance, the before mentioned neuter element can be observed in the context of a theory called algebra, but also in the context of a specific structure – groupoid (N, +). An arbitrary vector, for example $(a_1, a_2, \dots, a_n)$, $a_i \in R$, can be observed as an object in the context of a theory called linear algebra, but also in the context of a specific structure – n-dimensional vector space. See Drekalović (2015, 316-320).

object outside the theory *M*. We will be able, therefore, to explain the phenomenon described by the theory *T* without the theory *M*, which is in accordance with the intuition of dispensability. Eventually, it is obvious from the above given definition that we will state that a mathematical theory *M* is indispensable to the scientific theory *T* if and only if there is an *x* object of the theory *M* which is indispensable to the theory *T*, or is non-trivially dispensable but without a dispensable alternative which is not a part of the theory *M*.[21]

If we agree that the above definition describes to some extent the intuition of dispensability of a theory, let us examine from the formal standpoint the relation between a mathematical object and the theory which contains it. Obviously, dispensability of a mathematical theory and that of its object is not the same thing. It is far from that. To be more precise, according to D3, dispensability of a theory *M* entails dispensability of the objects of that theory. In other words, it cannot occur that a mathematical theory is dispensable to scientific explanations of a phenomenon and one of its objects is not, which trivially results from D3. On the other hand, on the basis of D3, generally, dispensability of an arbitrary object does not imply dispensability of the whole theory, with all its objects included. For example, some of the number theory objects, such as Euler's function, are dispensable to the cicada example according to D2, but that does not imply that the same goes when it comes to the entire theory. According to D3, as well as according to the expected intuition, number theory is indispensable to the mentioned phenomenon.

4. Conclusion

It seems that the EIA, in its present form, still cannot contribute to the strength of Platonism. This text has pointed to several reasons why that is the case. Firstly, the very formulation of the EIA contains elementary technical impreciseness related to the absence of appropriate quantifiers, which further

[21] To put it more formally, a mathematical theory *M* is indispensable to the theory *T* if and only if there is a mathematical object *x* of the theory *M* for which two following conditions are required:

1. *x* is not trivially dispensable;
2. *x* is not non-trivially dispensable or there is no a mathematical object *y* which does not belong to the theory *M*, and which is also non-trivially dispensable to the theory *T*.

extends the impreciseness onto the ontological level. This form of the EIA leaves one of the main questions about existence in mathematics unresolved. Namely, it is not entirely clear, as the EIA has shown, whether the Platonists aspire to discuss only the existence of a limited number of mathematical objects, without dealing too much with the objects whose existence could not have been granted or, contrary to that, the EIA has a significantly larger aim to fight for the existence of the ultimately defined object of all the mathematical theories.

Why should we expect a solution to this Platonist position exactly from the EIA? Is it not too much to expect that, as an argumentative tool of a very short history, it can be employed to resolve a question which has remained open from the very beginnings of Platonism? There exists at least one reason why the great hope is invested in this argument. With the EIA's modal and syllogistic formulation, it has already been indicated that there are unquestionable tendencies towards stricter and almost formal explanation of the existence of mathematical objects issue. That kind of logical explanation is, at minimum, expected to offer a completely clear proposition about its field of reference – only some or all the objects. This field of reference cannot be seen in the EIA.

Molinini has reminded recently that the lack of precision is a general deficiency of the Argument, pointing to the desirability of an additional definition of the dispensability concept, which is essential to the Argument. His contribution is important not only because of the efforts to define the concept of dispensability on the basic level, but also because he underlined that it makes sense to consider dispensability of a mathematical object only in the context of the entire mathematical theory to which the object belongs, and not as isolated and independent from other objects of the theory. However, as we have seen, those attempts have in a sense also displayed some of their own drawbacks, both formal and fundamental. They have also remained incomplete. By incompleteness we refer exclusively to the intuitive approach to the concept of dispensability of a mathematical theory, even though in his criticism of Baker, Molinini has started precisely with the idea that the intuitive notion about an object's (in)dispensability should be reinforced with somewhat more formal approach. There is no reason why dispensability of a theory should not acquire the same treatment. We have drawn attention to a technical shortage of the definition D1, the reason why it does not encompass some of the trivially dispensable objects, and we have then proposed the definition D2, which surpasses

that shortage. On the basis of that, as well as on the basis of the expected intuition, we have proposed the definition D3 of the dispensability of a theory. This has shown that (in)dispensability of a mathematical theory can by no means be the same thing as (in)dispensability of its arbitrary object.

References

ALKAUSKAS, G. (2009): A Curious Proof of Fermat's Little Theorem. *The American Mathematical Monthly* 116, No. 4, 362-364.

ANDRÉKA, H., MADARÁSZ, J. X. and NÉMETI, I. (2007): Logic of Spacetime and Relativity. In: Aiello, M., Pratt-Hartmann, I. and Benthem, J. (eds.): *Handbook of Spatial Logics*. New York: Springer, 607-711.

BAKER, A. (2003): The Indispensability Argument and Multiple Foundations for Mathematics. *Philosophical Quarterly* 53, No. 210, 49-67.

BAKER, A. (2005): Are There Genuine Mathematical Explanations of Physical Phenomena? *Mind* 114, 223-238.

BAKER, A. (2009): Mathematical Explanation in Science. *British Journal of Philosophy of Science* 60, 611-633.

BAKER, A. (2015): Parsimony and Inference to the Best Mathematical Explanation. *Synthese*, doi: 10.1007/s11229-015-0723-3.

BANGU, S. (2008): Inference to the Best Explanation and Mathematical Realism. *Synthese* 160, 13-20.

COLYVAN, M. (2001): *The Indispensability of Mathematics*. New York: Oxford University Press.

COLYVAN, M. (2002): Mathematics and Aesthetic Considerations in Science. *Mind* 111, 69-74.

DALY, C. and LANGFORD, S. (2009): Mathematical Explanation and Indispensability Arguments. *Philosophical Quarterly* 59, 641-658.

DODOS, P. (2010): *Banach Spaces and Descriptive Set Theory: Selected Topics*. New York: Springer.

DREKALOVIĆ, V. (2015): Some Aspects of Understanding Mathematical Reality: Existence, Platonism, Discovery. *Axiomathes* 25, No. 3, 313-333.

ERDOS, P. and SURANYI, J. (2003): *Topics in the Theory of Numbers*. USA: Springer.

GOLES, E., SCHULZ, O. and MARKUS, M. (2001): Prime Number Selection of Cycles in a Predator-Prey Model. *Complexity* 6, No. 4, 33-38.

LIGGINS, D. (2008): Quine, Putnam, and the 'Quine–Putnam' Indispensability Argument. *Erkenntnis* 68, 113-127.

LOOMIS, E. (1972): *The Pythagorean Proposition*. Washington: National Council of Teachers of Mathematics.

MADDY, P. (1990): *Realism in Mathematics*. New York: Oxford University Press.

MADDY, P. (1997): *Naturalism in Mathematics*. New York: Oxford University Press.
MELIA, J. (2000): Weaseling Away the Indispensability Argument. *Mind* 109, 455-479.
MELIA, J. (2002): Response to Colyvan. *Mind* 111, 75-79.
MOLININI, D. (2012): Learning from Euler. From Mathematical Practice to Mathematical Explanation. *Philosophia Scientiae* 16, No. 1, 105-127.
MOLININI, D. (2014): Evidence, Explanation and Enhanced Indispensability. *Synthese*, doi: 10.1007/s11229-014-0494-2.
PINCOCK, CH. (2012): *Mathematics and Scientific Representation*. Oxford: Oxford University Press.
PUTNAM, H. (1971): *Philosophy of Logic*. New York: Harper & Row.
SAATSI, J. (2011): The Enhanced Indispensability Argument: Representational Versus Explanatory Role of Mathematics in Science. *The British Journal for the Philosophy of Science* 62, No. 1, 143-154.
YABLO, S. (2000): Apriority and Existence. In: Boghossian, P. and Peacocke, Ch. (eds.): *New Essays on the A Priori*. Oxford: Clarendon Press, 197-228.

Organon F 23 (3) 2016: 299-323

Distribution and Inference: What Philosophical and Computational Semantics can Learn from Each Other

RADEK OCELÁK

Institute of Philosophy. Academy of Sciences of the Czech Republic
Jilská 1. 110 00 Praha 1. Czech Republic
radioc@seznam.cz

RECEIVED: 07-03-2016 • ACCEPTED: 26-04-2016

ABSTRACT: Distribution of a word across contexts has proved to be a very useful approximation of the word's meaning. This paper reflects on the recent attempts to enhance distributional (or vector space) semantics of words with meaning composition, in particular with Fregean compositionality. I discuss the nature and performance of distributional semantic representations and argue against the thesis that semantics is in some sense identical with distribution (which seems to be a strong assumption of the compositional efforts). I propose instead that distribution is merely a reflection of semantics, and a substantially imperfect one. That raises some doubts regarding the very idea of obtaining semantic representations for larger wholes (phrases, sentences) by combining the distributional representations of particular items. In any case, I reject the generally unquestioned assumption that formal semantics provides a good theory of semantic composition, which it would be desirable to combine with distributional semantics (as a theory that is highly successful on the lexical field). I suggest that a positive alternative to the strong reading of the distributional hypothesis can be seen in the philosophy of inferentialism with respect to language meaning. I argue that the spirit of inferentialism is reasonably compatible with the current practice of distributional semantics, and I discuss the motivations for as well as the obstacles in the way of implementing the philosophical position in a computational framework.

KEYWORDS: Lexical semantics – distribution – compositionality – inferentialism.

1. Introduction

One of the most crucial insights of the present-day computational, application-oriented approach to the semantics of natural language is this: we can usefully capture the meaning of a word by characterizing its distribution, or the contexts in which the word appears. As one famous aphorism goes, "you shall know a word by the company it keeps" (Firth 1957, 11). This proclamation may sound odd, and surely there are many ways of reading it. But it has been made clear by now that at least in some readings, the "distributional hypothesis" lends itself to remarkably successful computational applications. Models based on this insight have been applied to a variety of semantic tasks. Even if the results are still far from perfection, they generally seem to be far above anything achieved, first, in the other paradigms of semantic thinking, such as formal or cognitive semantics, and second, in the computational semantic branches that draw their inspiration from them.

Neither of these two points is quite surprising. As concerns the latter point, the distributional formulation of the natural language meaning problem is the key that enables us to treat the problem based on large amounts of actual language data, using the mechanical efficiency of a computer, or many computers at a time. It thus offers an interesting alternative to relying on our creative (see Schneider 1992) but relatively inefficient minds operating with language intuitions (which are, moreover, sometimes not too reliable). In the simplest case, word meanings as mysterious objects exclusively accessed by human minds are replaced by word meanings as patterns of textual co-occurrence of the target words with other words. Textual words being nothing but sequences of characters, that provides for efficient processing of the language material collected in extremely large corpora of written text. State-of-the-art models in distributional computational semantics are nowadays standardly built upon corpora containing billions of lexical tokens.

As concerns the former point, we might argue that the superior results in applications follow from the very nature of computational semantics, and computational linguistics in general. Computational linguistics differs from the theoretical approaches to language rather substantially in its orientation. At least as much as for theoretical *understanding* of language phenomena, the struggle here is for efficient "engineering" solutions to well-defined applied problems (such as machine translation or automatic summarization). Providing a good engineering solution nonetheless does not imply knowing *why* the

solution approximates the related phenomenon of natural language, or actually understanding that phenomenon. One might therefore claim that theoretical semanticists need not be unsettled by the success in application achieved by their computational colleagues.

In this paper, however, I would like to reverse that perspective in the following way. The tasks considered in the computational paradigm, and the distributional branch in particular, are not aimed to capture any sort of detached mechanical processes unrelated to human language use. Instead, they closely resemble some of the tasks that any competent speaker is likely to perform on a daily basis. Every now and then, we are expected to paraphrase, summarize, distinguish between two senses of a word, choose an appropriate synonym, sometimes even translate, etc. Suppose that a machine achieves human-like mastery in the *whole* spectrum of semantic tasks of which our everyday struggle with language consists. Then from a pragmatic point of view there will be little reason to claim that what the machine does has nothing to do with "real" semantics. Given the psychological and neurological aspects of our semantic competence, this machine will obviously not embody all there is to such a competence (or to the *implementation* thereof in our minds and brains). But it is also clear that the position that such an intelligent device has nothing whatsoever to teach us about "real" semantics would be absurd. Since the rise of automatic dishwashers, there have not been many complaints to the effect that what they actually do has nothing in common with *true* dish-washing as performed by humans.

Yet we are still nowhere near that ultimate stage in distributional computational semantics, and in the following I will try to argue that with a purely engineering approach we are not on our way there either. This is where theoretical understanding comes in. We should not claim that a machine ideally performing in semantic tasks would provide no such understanding to us. But it seems equally clear that without theoretical understanding of language we will not be able to bring a machine to such an ideal performance level (or any close to it) in the first place. Even if you occupy yourself with fairly practical tasks, you should not systematically ignore what appears as a good theory. Otherwise you might find yourself in the position of someone who keeps driving nails with a screwdriver, refusing all theoretical lessons in the mechanics of a hammer.

For this reason, I find it appropriate in this paper to combine two perspectives that are seemingly quite disparate in their assumptions and goals. First is

that of distributional semantics, as a very fruitful—although by no means exclusive—branch of computational semantics. Second is that of inferentialism, as a position in the philosophy of semantics. Interbreeding two remote perspectives, the paper of course runs the risk of not being digestible for either party. But I think that both approaches to natural language meaning can be mutually enhanced as to their *own* respective goals: success in semantic applications on one hand, understanding of meaning in language (and crucially, *validation* of such understanding) on the other. My hope is that the reader, having successfully navigated between the computational Scylla and the philosophical Charybdis, will be in a position to judge whether this claim is a correct one.

The paper is further structured as follows. In section 2, the stage is set by characterizing distributional semantics as to its basic ideas, methods, results and their broader significance. Further, some recent ideas regarding the possible enrichment of distributional semantics with semantic composition are discussed. (Except for the second part of 2.3, which is more critical in character, the whole section 2 is meant to be fairly consensual, and a reader who is familiar with distributional semantics and its recent development should feel free to just scan through it.) In section 3, I reflect on the theoretical status of distributional semantics, more specifically the question of the relation between distribution and meaning. I argue in favor of a weak, rather than strong, reading of the distributional hypothesis. In section 4, I return to the performance of distributional semantics from a more critical angle. With the observations made, I try to support the position that there is a serious gap between distribution and meaning, and I draw some consequences for the project of compositional distributional semantics. Finally, in section 5, I work towards presenting the inferentialist approach to semantics as a positive and viable alternative to the strong version of distributionalism.

2. Distributional semantics

2.1. Distributional semantic models

In this section, I outline the most important features of the distributional program in computational semantics. Note that this is just a very basic sketch. A much more thorough picture of the framework, its origins, assumptions, methods, goals and results can be found in works such as Lenci (2008), Turney and Pantel (2010), or Erk (2012).

Let me start the presentation with a toy example. Assume that the following table expresses how often each of the target words *dog*, *cat*, *tortoise*, *comb* occurred in the proximity of the words *hair* and *run* in a toy corpus.

	hair	*run*
dog	6	7
cat	8	6
tortoise	0	2
comb	5	0

Each row of the table determines a vector in a two-dimensional space, where each dimension corresponds to one of the context words; so, e.g., the vector for *cat* begins in the point [0,0] and ends in [8,6].

The distributional hypothesis generally states that the meaning of a word can be approximated by its pattern of occurrence in various contexts. Now, since the vector of each of the four target words is defined to (partly) capture just such a distributional pattern, we may decide to treat it as a *semantic representation* of the word in question. An important feature of vector semantic representations is that they are graded: a set of such representations is not merely a list of items (such as, for instance, the set of entries in a dictionary). We have a graded measure of similarity for any two of them: the angle formed by the two vectors in question, or more conveniently, the cosine of that angle. The smaller the angle (higher the cosine), the more semantic similarity we should expect between the words represented by these vectors. Thus in our toy example, at least some of the predictions will appear quite intuitive. (That is how the example is made up, of course.) *Cat* will come out as fairly similar in meaning to *dog*; *tortoise* not so much; *comb* will come out as particularly dissimilar from *tortoise*. One should note that no semantic information in any traditional sense went into these representations. All the table contains are (hypothetical but arguably plausible) co-occurrence counts of particular *words*.

There are literally dozens of reasons why the above does not constitute an adequate semantic analysis of the target words *dog*, *cat*, *tortoise*, and *comb*. However, a more interesting question is, which of the problems are—or can be—overcome by scaling the approach up with the available computing power, and by considering the many variants of the model that have been explored in distributional semantics up to now?

It is just for the sake of illustration that the previous example works with a small number of vectors in a two-dimensional space constituted by two context words, reflecting co-occurrence counts in a very small (hypothetical) corpus. In fact, the simple mathematics employed is easily generalized to multi-dimensional spaces with an arbitrary number of context dimensions. Thus the state-of-the-art distributional semantic models typically contain vectors for many thousands target words, vectors that "live" in several hundreds of dimensions. (Usually these are secondary dimensions which are gained from the original dimensions, given by many thousands of context words, by means of dimensionality reduction techniques.) As has been mentioned already, it is nowadays possible to build the vectors based on the co-occurrence counts in corpora of several billion textual words. That is, current distributional semantic models try to approximate lexical meaning using amounts of distributional information that are utterly incomparable to the toy example above.

Further, there are many alternatives to using the raw word co-occurrence counts as the basis of semantic representation. Some sort of automatic re-weighting of these counts is usual, or even necessary, so as to ensure that the more informative co-occurrences (such as that between *dog* and *bark*) will count more than those which are frequent but rather uninformative (e.g., *dog* and *the*). Also the notion of occurrence in a context can be made precise in various ways. Sometimes, it is defined as occurrence within a textual "window" of *n* word positions to the left and to the right from a particular token of the context word. Another option is to look for *any* co-occurrences within a single web-based document. It is possible to define the occurrence contexts in terms of lemmas[1] rather than plain word forms; or we can define the contexts with the use of syntactic characteristics (such as *dog* in the syntactic function of a direct object). The last two options depend on there being a method of automatic lemmatization or syntactic parsing applicable in the whole extent of the primary corpus, which is supposed to be as large as possible.

In theory (much less in practice, so far), *extralinguistic* contexts are considered as well. The fact that current models almost exclusively work with *textual* distribution seems to be a matter of contingent limitations rather than of

[1] Lemma is a representative form standing for the plurality of forms a lexical item can take, such as *bark* for *bark*, *barks*, *barked*, *barking*.

a theoretical commitment (cf. Lenci 2008, 10). Apparently, distributionalists are prepared to include as contexts whatever is technically manageable in a sufficiently large scale. For instance, some models derive their sets of contexts from large databases of labeled images. That seems important for the philosophical assessment of the program, for in this, distributionalism arguably diverges from the narrow, intralinguistic distributional analysis once practiced by the American linguistic descriptivism.[2] At the same time, it comes closer to the use-theoretic view of meaning originating from later Wittgenstein. After all, the hypothesis that the meaning of an expression is a matter of *where* it is used differs from the famous Wittgensteinian dictum solely by replacing *how* with *where*. That seems to open some room for a use-theoretic reappraisal of the distributional program, attempted in section 5.[3]

---

[2] Zellig Harris, the main descriptivist figure, is seen as a precursor of distributional semantics by Lenci (2008, 3ff.).

[3] It should be noted, finally, that there is also what Baroni et al. (2014a) call a new generation of distributional semantic models, represented notably by Mikolov et al. (2013). They are models that grew in the natural language processing field and the now dramatically developing area of neural network research, quite independently of the distributional tradition outlined above, which has more connections to theoretical linguistics. These models, referred to as *neural network language models* or *context-predicting models*, also semantically represent words with vectors in a multidimensional space. Instead of counting co-occurrences and applying heuristic transformations, however, the vectors are estimated by means of automatic learning, optimizing the success in the prediction of missing words in a known context. The evaluation by Baroni et al. (2014a) indicates, to the authors' own surprise, that these models perform consistently better than the traditional distributional models. In the following, context-predicting models are not systematically addressed. While I originally thought most of the critical considerations in this paper would apply to these models as well, Tomáš Musil (personal communication) pointed out to me an important difference which might prevent this from being the case. Namely, in context predicting models, the change in the semantic representation of an expression permeates further into the system by bearing on the representations of other expressions. That is not true in the traditional distributional models, where an expression's semantic representation is given by its co-occurence with other expressions but not by the representations of those expressions (which are, again, defined in terms of their own co-occurences).

2.2. The performance of distributional models

The previous technical characterization of distributional semantic models might appear omissible from the standpoint of some philosophical preconceptions about meaning which we may hold. But it is useful to see some details of the techniques that achieve as much in practical terms as distributional semantic models do. These models have been applied, with non-negligible success, to a variety of semantic tasks. From the theoretical perspective, many of these tasks are, in some form, part and parcel of our everyday operating with language. From the perspective of computational linguistics, methods successful in dealing with the tasks are likely to contribute to final language processing applications such as machine translation or question answering systems.

For instance, the performance of distributional models on the task of synonym detection is rather impressive, at least at first glance. The well-known TOEFL test consists of 80 multiple-choice questions where the subject is asked to pick one synonym for the target word out of four candidates (e.g., to choose the synonym *imposed* for the target *levied* from the candidate set *believed*, *imposed*, *correlated*, *requested*). In this test, the most successful distributional models, relying exclusively on the similarity of the vector representations of the words in question, are able to match in performance and even outperform the average college-educated native speaker of English (cf. Landauer and Dumais 1997; Baroni and Lenci 2010; Baroni et al. 2014a). Other tasks in which distributional models enjoy highly non-trivial success include, among others, prediction of human judgments of semantic similarity and relatedness, categorization of concepts into natural categories, detection of relational analogies (such as, *brother* is to *sister* as *grandson* is to *granddaughter*), even prediction of the psycholinguistic effect of semantic priming; (see, e.g., Erk 2016; Baroni and Lenci 2010; Baroni et al. 2014a; Baroni et al. 2014b; and their references.)

This is not to say that the current distributional models are able to solve all the semantic tasks that an average human speaker can, and with comparable accuracy. In fact, there is much that they *cannot* do in any satisfactory manner. (I will go into some detail in section 4.) But it is very much worth attention that they achieve relative success, and even approximate human performance, in *some*—undeniably semantic—tasks. This is especially manifest in comparison with the situation in formal semantics. In that field, there exists very little transparent evaluation in terms of what the proposed models can actually *do*,

which can be probably linked to the fact that they cannot do much in practical terms. (That seems to be agreed upon by the critics and the outsiders as well as the insiders of formal semantics, even if the other opinions regarding the value of formal semantic work differ; cf. Maddirala 2014.) By contrast, in computational semantics a lot of attention is traditionally paid to evaluation against independent data, and a substantial part of work goes into devising new evaluation methods, sets of testing data, etc.

Another difference from the more theoretical approaches to semantics, which is however closely related to the previous, is that distributional models require little[4] or no human "supervision", little or no semantic information brought in manually by semantically competent humans. They can thus be automatically trained for tens of thousands of target and context words on huge amounts of actual language data. This is not the case with formal semantic representations, which are typically crafted manually, as if one by one, by a semanticist, based on a small sample of actual language instances. (Here, I gloss over the fact that formal semantics hardly ever deals with problems of lexical meaning, whereas distributional semantics is, to a large extent, lexical semantics.) This is clearly an important part of the relative practical success of distributional semantics: with the limited descriptive capacities of individual humans, it is hard, or extremely expensive, to cover the vastness of human language use.

One more fact can be noted in favor of distributional vectors as genuine semantic representations in some sense, rather than as mere *ad hoc* engineering constructions. Although different parameter settings are often optimal for capturing different aspects of lexical meaning, one and the same distributional model can be used, with moderate success, for a plurality of purposes or semantic tasks. This thought is elaborated, e.g., in Baroni and Lenci (2010).

2.3. Composition in distributional semantics

An obvious drawback of the distributional approach to semantics as presented so far is the limitation to lexical meaning, or, in the best case, to the meaning of short and common phrases (such as *fall apart* or *kick the bucket*). Larger phrases and whole sentences will generally not occur in an arbitrarily

[4] Baroni et al. (2014a, 1): "Occasionally, some kind of indirect supervision is used: Several parameter settings are tried, and the best setting is chosen based on performance on a semantic task that has been selected for tuning."

large corpus with a frequency that could make the distributional information any informative in the semantic respect. (On the level of phrases and sentences, the number both of possible target vectors and of possible context dimensions grows tremendously, as presumably does the number of semantic distinctions that must be made. But there are not more tokens of phrases or sentences in a corpus than there are tokens of words, so the distributional information in the table of co-occurrence counts will be extremely sparse.) And indeed, semantic composition has recently been a hot topic in distributional semantics.

The question is: Can you combine the vector representations of particular words in a phrase (such as *black dog*) so as to obtain a useful semantic representation of that phrase, without having to rely on the distributional properties of the phrase as a whole? The most rudimentary attempts in this respect involve some very basic mathematical operations with the vectors, the resulting "phrasal" vector being obtained by simple addition or multiplication of the basic vectors. Some sort of linear weighting is possible, e.g., in order to stress the semantic role of nouns as compared to adjectives (Mitchell and Lapata 2010). These all are clearly very *ad hoc* solutions, with hardly any motivation other than mathematical simplicity.

A more ambitious program in compositional distributional semantics is formulated by Baroni et al. (2014b). Here, the idea of meaning composition as functional application, a fundamental notion from formal (model-theoretic) semantics, is adopted. Some words, nouns in particular, are represented in the familiar fashion, with their basic distributional vectors. Other words, such as adjectives, are semantically conceived as functions turning vectors into vectors; thus e.g. the vector for *black dog* can be obtained by the application of the functional meaning of *black* to the basic vector of *dog*. Yet other words are conceived as binary functions, etc., roughly in correspondence with the matching between grammatical categories and semantic types in Montague grammar (see e.g. Gamut 1991).

Despite the inspiration, this approach to semantic composition also differs from the formal semantic treatment in some important respects. First, unlike in formal semantics, the lexical functions are given concretely and informatively, not only defined as to their type and otherwise left unspecified (or specified just informally using disquotation, such as, "black" refers to the function that assigns truth value 1 to all black objects and only them). Namely, they are estimated based on the short phrases that still occur in the corpus often enough for their distributional representation to be semantically informative.

Basically, the functional representation of *black* is automatically estimated based on how the distributional vector of *black dog* differs from that of *dog*, that of *black book* from that of *book*, etc.[5] Once it is learned in this way, it can be used to derive the representations of longer phrases for which representation by the basic distributional vector cannot be assumed.

Second, the correspondence to the Montagovian matching between grammatical categories and semantic types is only partial, as attested by the treatment of common nouns such as *dog* (cf. Baroni et al. 2014b, 59). In formal semantics, common nouns, just like intransitive verbs or adjectives, are standardly conceived as logical predicates; that is, words with a functional meaning. The reason why Baroni and colleagues do not preserve this choice, in which the semantic types of nouns, adjectives and intransitive verbs are unified,[6] is clearly pragmatic. Representing common nouns with basic distributional vectors works remarkably well, and it would be unwise to force the distributionalist program into the scheme of formal semantics, a discipline whose outcomes are not nearly as efficient in practical terms.

But then, why should we bother incorporating *any* of the formal semantic tenets into the distributionalist program? It makes sense if we believe that formal semantics provides a good theory of semantic composition nevertheless. In any case, this is in accordance with how formal semanticists themselves tend to present the discipline (facing the lack of practical applications), and Baroni et al. (2014b) seem to share that belief. I do not, and I think there are serious reasons to believe the contrary. In Ocelák (manuscript), I attempt to elaborate these. Just briefly, my argument regarding formal semantics is that the lack of interest in lexical meaning, combined with the lack of empirical evaluation of the proposed semantic formulas, leads to the construction of chimerical compositional structures whose "adequacy" is a purely formal matter.

---

5 That is, the semantic representation of short phrases like *black dog* can be, in principle, either obtained by composing the representations of their parts, or specified directly as their basic distributional vectors. Given the method of estimating the functional representations, the outcome will typically be different in these two cases. The choice between the two options is upon the theorist. There is however also an argument for keeping both, pointing out the difference between the compositional and the idiomatic reading of, e.g., *kick the bucket* (Baroni et al. 2014b, 7).

6 That, in any case, is an option much more intuitive to logicians than to linguists.

For instance, the quantifier *all men* is in the most basic (extensional) case translated as $\lambda X \forall x(Man(x) \rightarrow X(x))$, which is supposed to be interpreted with a function that assigns truth values to functions from individuals to truth values (that is, to logical predicates). This function, however, is never given in full. It is only informally specified as *that function which* assigns the *appropriate* values to all relevant predicates (such as, 1 to *mortal* and 0 to *dark-haired*: for all men are mortal but not all of them are dark-haired). But that actually amounts to little more than saying that the meaning of a part is *whatever gives the right meaning* for the whole when applied to what we regard as the meaning of another part. It is then hard to see where such a quasi-analysis could possibly go wrong. At the same time, this can be found in the core of most formal semantic analyses. I therefore suspect that the existing body of work in compositional, lambda-phrased formal semantics can largely be seen as aprioristic elaboration of the Fregean idea of functional application. Whether the resulting theory of semantic composition is any good in empirical terms is highly questionable.

It moreover seems to me as a sort of wishful thinking to suggest that distributional and formal (or "denotational") semantics cover "complementary aspects of meaning" (Baroni 2014, 24; cf. also Erk 2016). The authors support this suggestion with the observation (in itself right) of the different focus in both approaches: generic knowledge in the former, episodic knowledge in the latter (Baroni 2014b, 22ff.). But at the same time, these approaches have been often pronounced complementary in dealing with the *lexical* and the *compositional* (or structural) aspects of meaning, respectively. How are these two divisions of labor supposed to square with one another? Surely, the distinction of the lexical and the compositional does not run parallel to that of the generic and the episodic. Lexical semantic competence, for instance, has both generic and episodic aspects to it. Thus the position that distributional semantics aims at the lexical *and* the generic, whereas formal semantics aims at the structural *and* the episodic, *and yet* they fully complement each other in the examination of language meaning seems problematic, even incoherent. For me, that as well constitutes a reason for being suspicious about the proposed boosting of distributional semantics with Fregean compositionality.

Altogether, I suggest we drop the assumption that formal semantics is a successful program in a domain that is complementary to the core domain of distributional semantics. And clearly, that would reduce the alleged need of encompassing both approaches in one framework.

As to distributional semantics alone, I have so far presented the framework in a more or less uncontroversial way, basically describing what people have done in the field. At this point, the very idea of enriching distributionalism with semantic composition invites a more philosophical discussion of the approach: an inspection of what it is that has been done, and what hopes we can (or cannot) derive from that.

3. What is distributional semantics, really?

Despite the general orientation on the performance in semantic tasks, the literature also contains explicit concerns about the philosophical interpretation of the distributionalist framework. In particular, people have made a distinction between a weak and a strong reading of the distributional hypothesis (see Lenci 2008, 14ff.; cf. also Baroni et al. 2014b, 20ff.).

Roughly speaking, distribution in the weak reading *reflects* the meaning of words (and perhaps also of some larger expressions), but does not *constitute* it. Words are generally used in accordance with what they mean (thus *dog* often appears in the context of *bark*, *bone*, *leash*, much less in the context of *fuel* or *oligarchy*). That makes distribution (which can be captured mechanically and efficiently) a useful guide in the exploration of meaning (which cannot), without however making it into a court of appeal as regards semantics. This conception leaves room for divergences of meaning and distribution, since it assumes that distribution is shaped also by factors other than meaning.

In the strong reading, distributionalism amounts to a cognitive hypothesis about the character of our semantic knowledge, or some parts of it. Here, vector space representations acquire the more binding character of cognitive or mental representations, rather than mere theoretical instruments. Sure, there is little reason to believe that the vectors we actually draw from a particular corpus, with a particular choice of target expressions, context dimensions, weighting techniques etc., capture the knowledge of any particular speaker very precisely. Thus distribution, at least as observable practically and in a large scale, can still somewhat diverge from meaning. But something like computing vectors based on the input and using them is (a part of) what is going on in our minds/brains when we acquire and use semantic knowledge—or so the thesis goes.

Baroni et al. (2014b), in their attempt to inject distributional semantics with compositionality, go for the strong reading of the distributional hypothesis. In opposition to them, I would like to defend the weak version of distributionalism here. By philosophical means, it is hard to disprove a cognitive hypothesis directly, stating facts by which it is contradicted. But I believe distributionalism can be presented in a way which will simply make the strong hypothesis not appear worth too much consideration.[7]

To me, it seems rather obvious that distribution is merely a reflection of semantics, and a substantially imperfect one. Apart from meaning, there are other important factors bearing on how words are put to use in a text; that is to say, factors that are also reflected in distribution. What the world is like is one of such factors. What we prefer to communicate about is another. (All these factors are interrelated and there are borderline phenomena: indeed much of the 20th century philosophy of language can be viewed as a struggle with the idea that they can be neatly separated and subsequently interlinked in a controlled fashion. But there are all sorts of clear cases which justify making the distinction nonetheless.)

Years ago, there was a fierce war in Bosnia, which made *Bosnia* co-occur with *war*, *tank* and *suffering* particularly often. Later, the situation stabilized, but people kept talking and writing about the past war. Yet neither of these periods added to the meaning of *Bosnia* a substantial something that we do not find in the meaning of *Switzerland*; neither made *Bosnia* markedly more related in meaning, e.g., to *war* than *Switzerland* is. I do not deny that many semantic changes do indeed proceed this way. But it is crucial to note that a semantic change is incomparably *slower* than the change in distribution to which it is linked. First, a massive change in distribution seems to be followed by hardly anything in the semantic respect. Slowly, something we call *connotation* may arise. It is only much later that a full-fledged semantic change can sometimes be recognized. Over past two centuries, *Waterloo* may have evolved into a synonym of *utter loss*, but very little of that change seems to have taken place in the first days or years after the co-occurrences of *Waterloo* in speech or writing rapidly changed in 1815. I believe this issue is overlooked

[7] That, incidentally, is a philosophical method of later Ludwig Wittgenstein, whom Lenci (2008) or Baroni et al. (2014b) mention among the historical sources of the distributionalist thinking.

when meanings are equated with distributional patterns, as seems to be more or less the case with the strong version of distributionalism.

Now, one can object that this, rather than being an objection to the strong reading of the distributional hypothesis, simply expresses a conservative view of meaning to which strong distributionalism provides a fresh alternative. Let me leave it at that for the moment: I hope to justify this conservatism later when a more positive program is finally outlined.

Provided that distribution is shaped also by factors other than meaning, its utility in the exploration of semantics may still be considerable, but is limited on principle. Consider an analogy: The ripples on Loch Ness may give us a clue about where underwater Nessie is at the moment. Yet the evidence is imperfect, since rippling is, besides the timid monster, also caused by the wind, by other creatures in the lake, etc. It would certainly be naive to insist that our methods of counting and measuring the ripples, and they alone, should make Nessie perfectly traceable, let alone to insist that the pattern of rippling is in some sense *identical* with her. To be sure, Nessie *can* be traced perfectly based on that pattern, but for that we would need to know the other factors and subtract their effects. By contrast, distributional semantics does *not* attempt to study the impact of factors other than semantics on distribution, and therefore is not in a position to subtract that impact.

4. Performance, nature and composition of distributional representations (again)

In section 2.2., I emphasized what distributional models are capable of doing in practical terms, in order to contrast them with other, more theoretical approaches to linguistic semantics. At this point, it seems convenient to mention what they have as yet *failed* to achieve.

Lenci (2008, 19ff.) identifies three main issuess with distributional semantics: semantic composition, reference or grounding, and inference. Of these, the first is discussed separately in this paper, and the second can perhaps be laid aside as a matter of technical limitations (see the discussion of extralinguistic contexts in section 2.1.). But the third problem, accounting for inferences, deserves some attention.

Inference, or entailment, plays a central role in a number of semantic approaches, including formal semantics and the inferentialist view of meaning

which is to be outlined in the next section. Correct inference, in the simplest case, is a transition between two sentences or utterances that is in a specific (namely, the *semantic*) sense appropriate.

It might seem that lexical semantics, the primary domain of distributionalism, does not concern sentential meaning at all, and therefore that we cannot expect this branch of semantics to provide an account of inference. That is however not quite true: the lexical semantic relations which are traditionally a crucial interest of lexical semantics are characteristic by licensing particular classes of inferences. Knowing that A is a *synonym* of B, we know that (by way of example and under certain additional conditions) we can infer "this is a B" from "this is an A" and the other way round. The information that A is a *hyponym* of B allows us to draw the inference from "this is an A" to "this is a B", but not the other way round. If A is an *antonym*, *meronym*, *co-hyponym* of B, that again seems to license at least some specific inferences in each case. Note that the same does not hold for the broad semantic similarity, which is supposed to be the relation primarily captured by distributional models. The information that A is semantically *similar* to B is not sufficient to license particular inferences from sentences containing A to sentences containing B.

Assuming there is a connection (to say the least) between understanding a sentence and knowing the appropriate inferences in which it is involved, it seems not unreasonable to expect of lexical semantics that it will do its part in accounting for inferences—that is, it will reliably detect lexical semantic relations. But for distributional semantics, with its basic notion of underspecified semantic similarity, this is a chronic problem.

It was mentioned above that the best of the current distributional models perform admirably on the standard TOEFL synonym detection task, easily reaching the performance of native human speakers. That is, however, a very specific task: it requires detecting exactly one synonym for a given term among three non-synonyms which also stand in no other particular semantic relation to the target. It is remarkable that this can be done very successfully on a distributional basis, but it is clearly not enough. In order to account for inferences, you need to be able to tell for arbitrary two terms whether or not they stand in the relation of synonymy, in the relation of hyper-/hyponymy, etc. A model's good performance in the TOEFL task does not guarantee this for synonymy. The vector representations of synonyms can be generally more similar to one another than those of semantically unrelated words, without the former being on the whole more similar than the vectors of antonyms, co-hyponyms etc.

And indeed, experimental results suggest that distributional models are too weak to tell apart cases of particular lexical relations reliably (Lin et al. 2003; Baroni et al. 2011.) Generally, the vectors most similar to the vector representation of a given word tend represent synonyms, co-hyponyms, and antonyms of the target word, without clear order. At the same time, not all synonyms, co-hyponyms etc. reach higher similarity than all words semantically less related to the target. That of course further complicates the classification task.

Admittedly, it is possible to construct the model or redefine the similarity measure so as to favor instances of a particular lexical relation; e.g., to enhance the "similarity" of co-hyponyms and suppress that of synonyms, antonyms etc. That seems to be the case at least for synonymy, co-hyponymy, and hyper-/hyponymy (cf. Baroni et al. 2011; Erk 2016). But the sorting success achieved is moderate in each case. For instance, one can find a specific similarity measure which, unlike the standard cosine measure, is likely to assign higher "similarity" on average to the instances of hyper-/hyponymy than to the instances of co-hyponymy (cf. Erk 2016, 21-22). That however does not imply that the measure is capable of sorting out hyper-/hyponymical pairs very efficiently. To give a parallel, men are no doubt taller than women on average; yet the utility of height alone in telling apart men from women is limited. The clue is better than random, but far from perfect. In accounting for inference, arguably, better than random is not good enough. You won't entrust a robot with making pancakes if its knowledge of appropriate inferences between sentences containing *egg*, *milk*, *food*, *poison*, *hot*, *cold* etc. is merely better than random.

As a side note, this approach also makes distributionalism as a cognitive hypothesis more problematic than it already seems to be. Namely, it is one thing to assume that what we do in our minds/brains when acquiring and using meanings is something like constructing and comparing distributional vectors. It is another thing, arguably a more involved one, to defend that we should actually need a whole bunch of vector spaces and/or similarity measures in order to cope with *various* lexical relations.

Above, the efficiency of distributional models in detecting lexical semantic relations is deliberately discussed in rather vague terms, despite there being many experimental results phrased in concrete numbers. I do not go into the evaluation numbers here, for that would make little sense in the absence of a detailed discussion of the respective semantic tasks, and of their relevance with respect to the problem in question. I nonetheless take it for given that the

current distributional semantic models, in spite of their achievements that are highly non-trivial from the point of view of theoretical semantics, are still far from giving a satisfactory account of lexical semantic relations (as an important part of natural language inference).

To this, we may react with the standard *more research is necessary* statement and keep trying to wring out what we can from distributional models. And no doubt, some improvement *can* be reached, in particular by exploiting ever bigger corpora and ever higher dimensionality, made possible by more efficient implementation and by using ever more computing power.[8] But my impression is that these improvements in performance are not promising enough to validate the position that in the limit, distribution *is* semantics.

Instead, I suggest that we bite the bullet of admitting that it is not. In my opinion, the problems with accounting for lexical relations are inherent to the approach as such. I believe that at the moment, the performance of distributional models is somewhere near the ceiling, and that is simply because distribution is a useful, yet imperfect reflection of semantics.

The hunt of Baroni and colleagues for composition in distributional semantics seems somewhat questionable from this perspective. In this view, composing distributional representations of particular words (even the advanced, functional representations) necessarily amounts to adding up the considerable imprecision that arises already on the lexical level. Very likely, there will still be some tasks on which the compositional representations (in particular those of relatively short phrases) will achieve a non-trivial performance. But if the claim that non-negligible amounts of error are being added up in composition is correct, then it is unclear whether such achievements can be of theoretical or practical consequence.

Let us go back to the Loch Ness parallel. If using a word's distributional pattern to explore its meaning is like tracing Nessie based on the momentary pattern of rippling, then the struggle for compositional distributional representations seems to be like trying to write up her biography based on a series of snapshots of the lake's surface. The former is limited in precision; in the latter, shortcomings are being piled up.

---

[8] Cf. Mikolov et al. (2013), who report on models which it took days on hundreds of processing cores to build up.

5. Distributionalism and inferentialism

I am aware that the previous critical considerations, pertinent as they may be, can hardly have much impact in lack of a positive alternative, one that would be viable from the point of view of computational linguistics. Also, one might want to bypass their theoretical relevance by insisting that the strong, cognitive distributional hypothesis gives rise to a radically new conception of meaning, whereby my assumptions regarding distribution, meaning etc. are simply not shared. But I think there *is* an alternative way to go, other than in the direction of contemporary compositional semantics. The alternative inspiration source is well-founded theoretically and I believe it can be stated precisely enough so as to invite computational implementation. Being use-theoretic in character, it seems to better fit the distributional reliance on language corpora, as documents of actual language *use*. There are moreover reasons to think that the implementation need not be quite disconnected from the current practice of distributional semantics.

I see such an alternative in the inferentialist philosophy of meaning, elaborated in particular by Brandom (1998); for a more accessible introduction, see Part I of Peregrin's (2014). Inferentialism draws on the idea that the meaning of a sentence is basically a matter of the appropriate inferences in which the sentence is involved. The meaning of a word, or generally of a subsentential expression, is then seen as its contribution to the inferential properties of the sentences in which it is contained. Here, the notion of inference is very broad, covering *language-language* transitions (that is, from sentences or sets thereof to sentences), as well as *world-language* and *language-world* transitions (that is, from worldly *circumstances* to sentences; and from sentences or sets thereof to worldly *actions*).

The inferentialist view is a specific elaboration of the Wittgensteinian idea that the meaning of a word consists in how the word is put to use, plus the aged observation that the primary use of a word is in the context of a sentence. It is specific, first, in that it emphasizes the normative character of our language use (the meaning of a sentence is identified not with its *actual* use, but with its *appropriate* use), and second, in that it narrows down the general notion of *use* to the transitions to which our sentences are subject. So, the meaning of a sentence (and closely related, the content of a belief) is given by what we *should* infer it from and by what we *should* infer from it in the context of other sentences (beliefs) which we are committed to assert (hold). Brandom's crucial

idea is that normative *statuses* of agents (i.e., what agents should do) can be reduced to factual normative *attitudes* (i.e., how the agents treat one another, as well as themselves, in relation to what they do). In this way, semantics is underlain by pragmatics. What people believe, or what their sentences mean, is explained—in a rather sophisticated way—in terms of what people do non-linguistically.

Argumentation for why Brandomian inferentialism is a fruitful and highly adequate philosophical approach to the semantics of natural language is far beyond the scope of this paper. Here, let me simply assume it is. On this assumption, I would like to make some comments towards bridging the gap between inferentialism as a philosophical project and distributionalism as a program in computational semantics, as I believe that enhancing a practical application with adequate philosophy is something desirable in principle.

The practical problem of inferentialism (which distributional semantics might be in a position to solve) is the following. Brandom's inferentialism is a holistic philosophy of meaning. What he draws is a picture of an overwhelmingly complex network in which any node standing for a sentence or a belief is deeply integrated. Any ordinary sentence is involved in myriads of appropriate inferences.[9] Little wonder that inferentialism as concerns natural language has not made it far beyond a mere philosophical idea until now: no *content* expression has ever been explicitly analyzed in inferential terms. Virtually the only inferentialist semantic analyses of natural expressions that seem plausible to some extent are the natural-deduction-style characterizations of sentential connectives such as *and*, *or*. ("A *and* B" can be appropriately inferred if A as well as B are given; from "A *and* B" we can appropriately infer A as well as B. That is all one needs to characterize the meaning of *and*, at least as traditionally employed in logic. The analysis is tempting in that we in this way completely avoid the need to postulate an object, typically a truth function, as the *meaning* to be mysteriously connected to the expression in question.) But

[9] Take, e.g., the belief/sentence stating that the cat is in the garden. It can be appropriately drawn from seeing the cat in the garden; or from hearing familiar noise from the garden; or from the belief that the cat was in the garden five minutes before plus the belief that it is an extremely lazy creature, etc. And given various "collateral commitments", it may be appropriate to infer that the cat is safe from the street traffic, or that there will soon be no mice in the garden, or that the cat will make a mess when it's back in the house; or it can lead to a *lemme-drive-the-cat-out-of-the-garden* practical commitment, etc.

application to this restricted vocabulary can hardly provide sufficient validation for such a general philosophical theory.

Think as we may that inferentialism is the right way of thinking about meaning, it cannot be considered an option by computational semanticists unless it is presented as viable by their methods. Preferably, it should be made feasible using the valuable resources that are available and that make computational linguistics successful as it practically is: large corpora of actual language use in the first place. In my opinion, inferentialists should side with the idea of computational implementation of their program. At least, the philosophical ambition of inferentialism is to reduce the mysterious notion of meaning to something more transparent, something that we *do*: something that computers, therefore, might be also capable of doing one day.

Here is why I think inferentialism is fundamentally compatible with the distributional perspective. Recall that distributional semantics attempts to capture the meaning of an expression in terms of its occurrence *in various contexts*. Usually, these are lexical contexts, so what is typically counted are lexical co-occurrences. But the distributional project does not set any *a priori* bounds to what we can regard as contexts. Various options have been considered: among others, lexico-syntactic contexts, web-based documents, extralinguistic contexts (such as labeled images)—and crucially, we may think of *inferential contexts* as well. We may want to count a sentence's occurrences in the context of sentences inferred from it, and in the context of sentences from which it is inferred.

There is a number of problems with this proposal. The first is that what primarily features in an inference are *sentences.* As mentioned in section 2.3, the actual co-occurrence information in the co-occurrence space of sentences (unlike the space of words) is extremely sparse for corpora of all available sizes—sparse beyond imagination. We could count co-occurrences of *words* in inferential contexts, but it seems to be of little use to know, e.g., that *freezing* and *green* co-occured within the inference "It is freezing outside. – I'd better take the green cap, the wooly one, since the red is really thin." (The co-occurrence of *freezing* and *wooly*, or *outside* and *cap* is perhaps more informative, but it occurs to me that counting word co-ocurrences would open the door for the same kind of imprecision that has been observed in the standard distributional representations.) Somehow, we need to treat a sentence as a whole, nevertheless. I do not have a solution for this, I only hope one can be given. Perhaps, a clever engineering solution can exploit the idea that the meaning of

a word is the way it contributes to the inferential properties of the sentences in which it is involved, and perhaps, the process of inferential characterizing can be bootstrapped from minimal inferences such as "this is a banana: this is yellow". Syntactic information will be surely indispensable in such a scheme.

Second, the issue with *world-language* and *language-world* inferences. Given the technical difficulties limiting the utilization of extralinguistic contexts, I suggest that we follow current distributional semantics in focusing on linguistic contexts, at least for the time being. That is, we may focus on *language-language* transitions. (Existing distributional models have shown clearly enough that non-trivial practical success can be hoped for even in the absence of extralinguistic information.) One more thing that needs to be technically overcome is that often, *language-language* inferences are inferences not from individual sentences, but from *sets* of sentences.

Third, the problem of normativity. Bradomian inferentialism explains the meaning of a sentence in terms of appropriate inferences, not in terms of actual inferences. Contrariwise, what we can (at best) gather from a corpus of actual language use are the inferences people make, possibly the inferences they make *regularly*, but not the inferences they *should* make. Here again, I suggest we take a pragmatic stance. The practical success of distributional models (that is, on tasks that are unequivocally semantic in nature) indicates that the cleft between actual and appropriate use is narrow enough for at least some practical purposes. One may here also consider Davidsonian arguments to the effect that it is incoherent to assume a massive amount of factual or semantic error among speakers (cf. Davidson 1974).

The fourth problem is likely the most serious one in practical terms. Actual inferences occurring in a corpus are not very reliably marked with formal means such as *therefore*, *thus*, *so*, etc. Yet worse, rudimentary inferences such as "this is a banana: this is yellow" scarcely make it to the communication of competent speakers. Usually, such inferences are assumed rather than pronounced. What gets explicitly communicated instead are complex inferences relying on a number of collateral commitments or shared assumptions: "People still remember the Denver incident. *Therefore*, Smith won't get more than 15 percent of the votes."

An option that comes to mind in this context is utilizing language *acquisition* corpora, rather than corpora of grown-up communication. Unfortunately, the corpora of the former type are several orders of magnitude smaller in size, which may be hard to bite for a distributionalist, and the data is very expensive

to gather. Its quality could nonetheless make up for that. It is first and foremost with children that we explicitly state what is otherwise obvious, talking about the color of bananas, etc.

This approach also seems to constitute an additional answer to the normativity issue. In talking to children, we are generally engaged not only in communication, but also in tuition and training that are relevant for the child's future communication. Thus stating "This is a banana. (*So*) it is yellow" in this situation is not merely an actual inference. Much more it amounts to the formulation of an inferential rule, to stating what inferences should be drawn. [10] Even so, there remains the problem that not all inferences are formally marked, and an amount of manual annotation may be necessary.

6. Conclusion

No doubt, the difficulties involved are considerable, and the "inferentialized distributionalism" just proposed may not reach the practical performance of the current distributional models any time soon. Still, I believe something in these lines is worth elaborating. Distributionalism in computational semantics has had highly non-trivial achievements, but in the end that all comes down to clever exploitation of the fact that meaning is reflected in distribution. If that is not *all* there is to meaning, the prospects of exploiting the idea further are of course limited.

Ultimately, the goals of computational and philosophical semantics cannot be as divergent as they possibly appear to be at the moment. Computational semantics is supposed to come up with something that can do what natural language meaning does, or what humans do using their semantic knowledge. Fair enough, but why would we think this can be achieved without paying attention to our best opinions about what natural language meaning *is*?

What there is in the project for inferentialism as a philosophical program seems also quite clear. Boosting computational semantics with inferentialist insights would constitute important empirical validation for the philosophical

[10] Note that there would be no point in stating such rules incorrectly. Joking or lying about bananas makes sense only after the child has mastered some basic inferential properties of banana-related sentences.

theory. A theory of an empirical phenomenon, as human language altogether is, has surely no right to spurn such a prospect.

Acknowledgments

I am grateful to Jarda Peregrin and Tomáš Musil as well as to several anonymous reviewers, for their relevant comments on previous versions. All remaining flaws and omissions are my own responsibility. Work on this paper was supported by the Research Grant No. 13-21076S of the Czech Science Foundation.

References

BARONI, M. and LENCI, A. (2010): Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36, No. 4, 673-721.

BARONI, M. and LENCI, A. (2011): How We BLESSed Distributional Semantic Evaluation. *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 1-10.

BARONI, M., DINU, G. and KRUSZEWSKI, G. (2014a): Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238-247.

BARONI, M., BERNARDI, R. and ZAMPARELLI, R. (2014b): Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in Language Technology* 9, 5-110.

BRANDOM, R. (1998): *Making It Explicit: Reasoning, Representing and Discursive Commitment*. Cambridge, (Mass.): Harvard University Press.

DAVIDSON, D. (1974): Belief and the Basis of Meaning. *Synthese* 27, 309-323.

ERK, K. (2012): Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass* 6, No. 10, 635-653.

ERK, K. (2016): What Do You Know About an Alligator When You Know the Company It Keeps? *Semantics and Pragmatics* 9.

FIRTH, J. R. (1957): *Papers in Linguistics*. London: Oxford University Press.

GAMUT, L. T. F. (1991): *Logic, Language, and Meaning. Vol. 2: Intensional Logic and Logical Grammar*. University of Chicago Press.

LANDAUER, T. K. and DUMAIS, S. T. (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104, No. 2, 211-240.

LENCI, A. (2008): Distributional Semantics in Linguistic and Cognitive Research. *Italian Journal of Linguistics* 20, No. 1, 1-31.

LIN, D., ZHAO, S., QIN, L. and ZHOU, M. (2003): Identifying Synonyms Among Distributionally Similar Words. *Proceedings of the 18th International Joint Conference On Artificial Intelligence*, 1492-1493.

MADDIRALA, N. (2014): *Philosophy of Logical Practice: A Case Study in Formal Semantics*. Master thesis, ILLC, University of Amsterdam.

MIKOLOV, T., CHEN, K., CORRADO, G. and DEAN, J. (2013): Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

MITCHELL, J. and LAPATA, M. (2010): Composition in Distributional Models of Semantics. *Cognitive Science* 34, 1388-1429.

OCELÁK, R. (manuscript): Besieging Model-Theoretic Semantics. Available at: http://ocelak.cz.

PEREGRIN, J. (2014): *Inferentialism: Why Rules Matter*. Basingstoke, UK: Palgrave Macmillan.

SCHNEIDER, H. J. (1992): *Phantasie und Kalkül: Über die Polarität von Handlung und Struktur in der Sprache*. Berlin: Suhrkamp.

TURNEY, P. D. and PANTEL, P. (2010): From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Organon F 23 (3) 2016: 324-366

Non-cooperative Strategies of Players in the Loebner Contest

PAWEŁ ŁUPKOWSKI

Department of Logic and Cognitive Science. Institute of Psychology. Adam Mickiewicz University
Szamarzewskiego 89a. 60-586 Poznań. Poland
pawel.lupkowski@amu.edu.pl

ALEKSANDRA RYBACKA

Department of Logic and Cognitive Science. Institute of Psychology. Adam Mickiewicz University
Szamarzewskiego 89a. 60-586 Poznań. Poland
rybacka.ola@gmail.com

RECEIVED: 14-04-2016 • ACCEPTED: 20-06-2016

ABSTRACT: In this paper the idea of the Loebner contest as a practical implementation of the Turing test is presented. The Brian Plüss' measure of the degrees of non-cooperation in a dialogue is applied to the dialogues of the Loebner contest. The proposal of a typology of non-cooperative features in the contest's dialogues is discussed and the reliability of annotation with the use of this typology of features is analyzed. The degrees of non-cooperation of judges and programs for the Loebner contest (editions 2009 – 2012) are presented and discussed. On the basis of the results the role of a judge and the strategies used by programs are discussed for the contest and the Turing test.

KEYWORDS: Turing test – Loebner contest – strategy – non-cooperation degree measure.

0. Introduction

The Turing test is widely discussed by philosophers, psychologists, computer scientists and cognitive scientists (see, e.g., Konar 2000; Harnish 2002).

Although it was proposed more than fifty years ago, the Turing test is still considered as an attractive and fruitful idea, when it comes to its theoretical aspect (see Saygin et al. 2001; Shieber 2004; Epstein et al. 2009) as well as its practical applications (e.g. the Loebner contest or CAPTCHA systems[1]). The main aim of this paper is to establish and analyze the measures and structures of non-cooperative verbal behaviors in the Loebner contest, which is the best known practical implementation of the Turing test. We have decided to analyze the Loebner contest conversations because they constitute a useful and reliable data source. This is a result of several factors. Firstly, the contest has been held yearly (since 1991) and its conversation logs are available publicly to researchers. Secondly, the core rules of the contest are the same every year and they stem from Turing's ideas. What is more, the conversation logs are supplemented with additional information, including judged scores and time-stamps. Last but not least, judges often ask the same question simultaneously to a program and to a human participant – this gives an opportunity to study the differences and similarities of the provided answers. In our opinion, the study of the Loebner contest may be beneficial in many fields, from testing Turing's original ideas concerning the test (when Turing proposed his famous test he came up with certain predictions about possible algorithms and behaviors for the test situation) to the practical results and clues about the Loebner contest setting (e.g. in identifying useful strategies for program players and for judges in the contest). What is more, this study can contribute to better design of contests based on Turing's ideas.

The motivation for our research is twofold. On the one hand, we may point at formal analysis of the Turing test setting presented in Łupkowski (2011) and Łupkowski and Wiśniewski (2011). On the other hand, our work is motivated by recent analysis of practical implementations of the Turing test (see e.g. Epstein et al. 2009; Łupkowski 2013; Warwick and Shah 2015; 2016).

The paper is structured as follows. In the first section, we briefly describe the Turing test (hereafter TT) idea and the rules and the setting of the Loebner contest (LC). We also introduce two issues that are often discussed in the context of TT, namely the role of a judge in the test and the issue of strategies that

[1] CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart. The main task of a CAPTCHA is to differentiate bots and human users in on-line services; see Ahn et al. (2003).

should be used by programs. These issues will be later discussed in the context of results from our study. In the second section, we introduce the concept of non-cooperation in a dialogue and its measure proposed by Brian Plüss (see Plüss 2009; 2010; and Plüss et al. 2011). We describe the set of non-cooperative verbal behaviors for LC that we use in our study. The third section contains the description of our main study in terms of the study sample, the method used, obtained results and discussion of their reliability. We end up with the summary and discussion of the issues introduced in the first section in the light of the study output.

1. The Turing test and the Loebner contest

1.1. The Turing test

The setting for the test proposed by Turing[2] might be presented as follows: the *interrogator*, and tested agents: a human and a machine take part in the test.[3] Parties of the game cannot see or hear each other, communication goes through written messages. It is the interrogator who asks questions and the players answer them (players are not permitted to ask any questions) – cf. Newman et al. (1952, 4). As for the questions' subject area, Turing seems to leave a free hand for the interrogator (cf. Newman et al. 1952, 5; Turing 1950, 434-435). Types of questions, as well as topics should not be restricted, and the conversation should resemble those in real life. As Turing puts it:

---

[2] We rely on the following sources in which Turing writes or speaks about the test: "Intelligent Machinery" (Turing 1948), "Computing Machinery and Intelligence" (Turing 1950), "Can Digital Computers Think" (Newman et al. 1952), "Intelligent Machinery, a Heretical Theory" (Turing 1951), "Can Automatic Calculating Machines be Said to Think?" (Newman et al. 1952), and "Digital Computers Applied to Games" (Turing 1953). For an overview of the discussion on TT rules see e.g. Saygin et al. (2001), Copeland and Proudfoot (2009), Łupkowski (2011) and Łupkowski and Wiśniewski (2011).

[3] The test with only two participants, interrogator and a tested agent (computer or human), is also often considered under the name *viva voce*. For an overview of terminology used in the context of TT see Harnish (2002, 183).

> The questions don't really have to be questions, any more than questions in a law court are really questions. […] 'I put it to you that you are only pretending to be a man' would be quite in order. (Newman et al. 1952, 5)

The role of the interrogator is to identify which of the players is a human and which is a machine only on the basis of collected answers. The interrogator wins a game when he/she makes an accurate identification. Otherwise, the interrogator loses the game.

1.2. The Loebner contest

The contest takes the name from its founder – Hugh Loebner. LC identifies the program with the best scores as the winner, and its programmers are awarded an annual cash prize. The winner does not need to be recognized as a human, but it has to be the most human-like among the other machine participants. The first computer program to pass the Turing test will be awarded a grand prize of $100,000.[4]

The design of the Loebner contest is meant to resemble Turing's proposal as closely as possible. However, the contest initially differed from Turing's original assumptions. In the first competition (in 1991) six programs and four people were accepted as participants, and ten judges were selected from respondents to a newspaper advertisement. The capability of computers at that time was insufficient to pass an unrestricted test, so the topic of conversation was limited and judges were asked to refrain from "trickery or guile". Restricting topics led to several problems. In 1992, the topic was hockey, and the lack of hockey fans among the judges led to more difficult and unusual questions (cf. Mauldin 1994). Hugh Loebner pointed out other problems with topic restriction, such as unnecessary complexity, a lack of fluency in dialogues and having to decide if the conversation stays on topic. Loebner proposed no restrictions on language used (allowing also for vulgarity or obscenity) and also no restriction on sensory modalities and the possible participation of robots in the future (see Loebner 2009). The contest has been unrestricted in the mentioned aspects since 1995.

---

[4] See the Loebner contest homepage: http://www.loebner.net/Prizef/loebner-prize.html.

The rules changed throughout the years, with the number of participants getting smaller, down to four computer programs, four human participants and four judges. We may sum up the core LC contest rules in the following way:

1. Before the final contest there is a preliminary phase aimed at choosing four best programs.
2. 4 human players, 4 AI players and 4 judges take part in the contest.
3. Each of the judges conducts simultaneous, split-screen conversations with two players without knowing their identity. One of the players is always a computer program and the second one is human. One such conversation is called the round.
4. In four rounds each player has a conversation with each judge.
5. Topics of conversations are unrestricted.
6. At the end of each round each judge will declare one of the two entities to be the human.
7. At the end of the contest the judges rank programs from the most human to the least human and assign points – the lower the score, the better.

In LC a judge holds a conversation with two participants, a human and a program, in each round. What is important, a judge knows that one of the participant is a computer program. Data is transmitted character by character, so that the opponent sees the typing process in real time. That requires a machine to imitate human speed of typing, as well as spelling mistakes. Loebner developed his own standard for a communication program to enable interaction between the participants during the contest. Since technology and the Internet become more and more prevalent, there are various ways for a computer program to interact with the world. Robby Garner proposed the standard interface for the Turing test, called *The Turing Hub* (see Garner 2009). Tests of this solution showed that programs running *via* The Turing Hub receive better scores in LC. This is due to the fact, that the hub eliminates visual clues, like typing and delays. In Gardner's opinion, the contest should be based strictly on verbal outputs and not on imitating the whole spectrum of human behavior.

The Loebner contest has well established rules and is held every year, and what is the most important, transcripts from each year are available for analysis. The Loebner contest is designed to implement Turing's original idea as accurately as possible. Therefore, it provides an interesting source when one wants to analyze some of Turing's assumptions, such as the one saying that the program should not reveal its identity. The dialogical form of the contest is perfect for analyses of participation in dialogues, both in terms of artificial intelligence studies and human linguistics. Organizers of the competition provide data and transcripts from each edition, containing information such as judges' names, scores they have assigned to participants, and pragmatic dialogue information like the time-stamp of every character. A LC conversation can be replayed in real time by using the program called *the Loebner Player*.[5]

1.3. Important issues of the test/contest situation

As we have mentioned in the Introduction, there are two issues of TT that are also reflected in LC. These are: (i) the program participants' strategies and (ii) the role of a judge in the test situation.

According to Turing, a computer should follow certain rules in order to win a game, that is, trying to behave like a human being as much as possible, including writing slowly, making spelling mistakes, hesitating before answering, and similar techniques. Turing says:

> The machine would be permitted all sorts of tricks to appear more man-like, such as waiting a bit before giving the answer, or making spelling mistakes. (Newman et al. 1952, 5)

However, we may imagine situations when a program will reveal its identity during the conversation. Will this affect the score in practical implementation of the test? We may also imagine another situation, namely a human being pretending to be a program. There are no rules in TT or LC that prevent such a behavior. When we think about LC also another possible question arises – namely, is this issue important in the light of contemporary programs'

---

5 See http://www.loebner.net/Prizef/loebner-prize.html.

performance? In other words, are modern dialogue programs taking part in LC sophisticated enough to successfully implement Turing's advice?

The second issue is related to the interrogator's perspective in the TT. This is one of the central issues when we think about evaluating this test setting (see Łupkowski and Wiśniewski 2011). We may consider two sub-problems in this area: the first one is how to select the interrogator to take part in the TT; the second one is how should the interrogator run the test.

The first problem has been widely discussed in the literature. Alan Turing's suggestion is that the interrogator should be a person who is not an expert in the field of computing machines (cf. Turing 1950, 442; Newman et al. 1952, 4). This restriction comes from the fact that Turing was aware that beliefs and knowledge of the interrogator may play an important role in the way of running the test. This issue is sometimes seen as one of the main drawbacks of TT. Exemplary argumentation might be the one formulated by Ned Block. He writes:

> Construed as a proposal about how to make the concept of intelligence precise, there is a gap in Turing's proposal: we are not told how the judge is to be chosen. A judge who was a leading authority on genuinely intelligent machines might know how to tell them apart from people. For example, the expert may know that current intelligent machines get certain problems right that people get wrong. […] A stupid judge, or one who has had no contact with technology, might think that a radio was intelligent. People who are naive about computers are amazingly easy to fool […]. (Block 1995, 379)

To sum up, according to Block, judges are easily fooled by well designed, but not intelligent computer programs. At the same time, they are more likely to reject a genuinely intelligent machine that has not mastered conversation skills. The problem of selecting an interrogator for TT becomes even more important when we think of the Loebner contest (and of any other implementation of the test). In such a case, the outcome of a dialogue is determined to a large extent by the judges. LC is a competition, and as such it should be governed by strict rules and regulations – including the one, which will determine, how to choose the interrogator (judge). There are many detailed proposals for this issue, however it is far from being solved. Loebner (2009) recommends journalists as the best judges. He claims they are willing, intelligent and, which may be the most important factor, have the power of publicity. On

the other hand, Garner (2009) disagrees with that opinion, suggesting that the selection of judges should be representative of the general population.[6]

The second part of the discussed issue received less attention in the literature. Let us remind the reader that in LC a judge is aware that he/she holds a conversation with two participants one of which is a computer program. Will this affect the LC conversations? Zdenek (2001) suggests that in such a test situation judges will behave like interrogators. They understand their task as revealing the true identity of the interlocutor as quickly as possible – treating LC as a kind of win/lose game. They start a conversation presuming that they are talking to a machine and change their mind only after this is proved to be otherwise. This kind of approach may influence a conversation, resulting in a series of questions instead of a regular chat and, supposedly, in many non-cooperative behaviors of the judges.

2. Measuring non-cooperation in dialogue

Many studies focus on types of interactions which are cooperative, where participants in the conversation have a common goal and are interested in achieving it effectively (think of the cooperation principle by Grice 1975). In this paper we are more interested in the situations where individual goals of dialogue participants are in conflict with their discourse obligations – this leads to non-cooperative verbal behaviors in a dialogue (cf. Plüss et al. 2011, 213). We may observe such behaviors in everyday conversations. They are however even more clearly visible and characteristic for certain types of dialogues, such as: interviews, interrogations and exams, where the goals of participants can differ and therefore more cases of deliberate non-cooperativeness emerge. LC resembles interrogation in its nature. It might be also described as a game, in which the goal of the judge is to tell the machine and a human apart. Thinking of it in this way, one may expect that many non-cooperative behaviors will occur on the part of the judge, as he/she will try to reveal the opponent's identity as quickly and effectively as possible. Current technology is advanced

[6] In this context the *Minimal Intelligent Signal Test* (MIST) designed by McKinstry (1997) is worth mentioning. The idea of MIST is to solve the judge issues in TT by making the judging process easy and possibly automatic. This supposed to be obtained by using a set of yes/no questions only.

enough to create a truly human-like dialogue program, and thus conversation between a robot and a human will result in many non-cooperative strategies, like changing the topic or refusing to answer questions. By measuring how often non-cooperative behaviors occur in LC, we aim at better understanding of strategies of players in this contest as well as the impact of its setting (described in Section 1.2) on the outcomes of the contest.

2.1. Brian Plüss' measure of the degrees of non-cooperation in dialogue

Brian Plüss focuses in his studies on political debates (see Plüss 2009). The reason is that these are the types of conversation that are highly non-cooperative in the sense explicated above. What is more, in this case non-cooperation is not a result of incompetence but is rather a rational strategy. As he points out, in the United Kingdom, journalists have a very incisive approach to political candidates, and at the same time politicians are trained to avoid subjects that are not favorable to their image, while focusing on delivering key messages to the public.

The degree of non-cooperation (DNC) proposed by Plüss is a measure that indicates how often interlocutors do something that leads to a break in the natural flow of conversation. In the case of the Loebner contest, we examine the verbal behaviors that are semantically non-cooperative or are in conflict with the rules of the contest.

The idea is to annotate dialogues using a certain set of *non-cooperative features* (NCF) which is characteristic for a given dialogue type. The ratio between the number of occurrences of NCFs and the total number of utterances is the degree of non-cooperation (DNC). The first part of the procedure is to establish a set of NCFs which are characteristic for a given dialogue type. Plüss proposed a list of such features for political debates and grouped them in three categories: (i) turn-taking, (ii) grounding and (iii) speech acts, later to be abbreviated to 5 basic *non-cooperative features*:

O	overlap;
GF	grounding failure;
UC	unsolicited comment;
I	interruption;
TC	topic change.

These features can be observed in the example of the tagged part of the interview between the BBC presenter Jeremy Paxman (P) and a former Home Secretary Michael Howard (H) (see Plüss 2010, 1):

P: (overlapping) Did you threaten to overrule him? (O)

H: … Mr. Marriot was not suspended. (GF)

P: Did you threaten to overrule him? (GF)

H: (pauses) I have accounted for my decision to dismiss Derek Lewis…

P: (overlapping) Did you threaten to overrule him? (O)

H: …in great detail before the House of Commons. (UC)

P: I note that you're not answering the question whether you threatened to overrule him.

H: Well, the important aspect of this which it's very clear to bear in mind… (GF)

P: (interrupting) I'm sorry, I'm going to be frightfully rude but… (I)

The brief (simplified) summary of Plüss' procedure is the following:

1. Establish set of *non-cooperative features* (NCF).
2. Annotate utterances using NCF categories.
3. Count *degree of non-cooperation* (DNC) for the dialogue.

Plüss' studies provide a better understanding of the nature of political interviews. They may be a useful tool to improve public debate and point out the possible effects of non-cooperation. His motivation was to construct a computational model of non-cooperative dialogues and to develop a system that deals with them. Research on non-cooperative speech behavior leads to better understanding of the dialogue structure and pragmatics and in general results in new ways of coping with a wider range of verbal behaviors.

2.2. The method of establishing DNC for the Loebner Contest

In our opinion the approach proposed by Plüss – with slight modifications – may be applied in the study of conversations in the Loebner contest. First of all, this contest resembles an on-line chat more than a natural face-to-face conversation. The flow of conversation is limited by the interface. No visual or auditory

cues are present, the dialogue is divided into utterances, which makes it relatively easy to notice any disturbances. Secondly, both political debate and the Loebner contest have rules for participants to follow. As Heritage states:

> If interviewers restrict themselves to asking questions, then they cannot—at least overtly—express opinions, or argue with, debate or criticize the interviewees' positions nor, conversely, agree with, support or defend them. (Heritage 1998, 8)

This corresponds with the Loebner contest's rule that forbids judges to express personal opinion during a conversation. Further on we read:

> Correspondingly, if interviewees restrict themselves to answers (or responses) to questions, then they cannot ask questions, nor make unsolicited comments on previous remarks, initiate changes of topic, or divert the discussion into criticisms of the interviewers or the broadcasting organization. (Heritage 1998, 8)

In the original setting, the Turing test is a kind of interview, where judges ask questions and players only answer them. In practice (as we may observe in the Loebner contest) conversations are more casual, with players often asking questions or changing topics.

On the other hand, TT also has many game-like features; participants have contradictory goals: a judge is supposed to tell a human and a program apart, and the program's task is to deceive the judge. Because of its competitive nature many examples of non-cooperative behavior are present on both sides. The computer program, just like a politician, is supposed to avoid topics that are not well established and can reveal the program's true identity, so it has to use tricks and indirectly guide the discourse in the right direction. The judge, just like an interviewer, has to ask the right questions and reveal the identity of an interlocutor.

Lastly, the use of NCFs seems to be adequate in LC context, because we focus on studying non-cooperative verbal behaviors on a semantic level. In dialogues between a man and a machine, it seems difficult to attribute intentions (and thus pragmatic level) to a machine. Unfortunately, judges usually discover the identity of a program after a couple of sentences, immediately changing their approach. They can have trouble assigning intentionality to a machine, thus they do not involve the cooperative principle in the sense presented by Grice. In such a situation NCF method still may be used effectively.

A basic set of NCFs for studying non-cooperativeness in LC was proposed in Łupkowski (2013). Four non-cooperative features were introduced and used to investigate the 2010 LC edition conversations:

IR a judge or a participant fail to provide a relevant contribution;

TC topic change;

RD request drop;

PO a judge expresses opinion about the player's performance.

For the study presented in the present work we have decided to add two more NCFs, to cover more types of behavior found in transcripts. The final set of NCFs used for the study (supplemented with explanations and examples) is presented below. Note that the original spelling is preserved in all the examples. Spelling errors are commonly used by judges and by programs in the LC.

IR (*fails to give a relevant answer/ask a relevant question*). The dialogue participant's response has no relation to the previous statement in a conversation. The player (P) asks a question that has no connection with the previous statement or is irrelevant. With this category we will also address situations, in which it is clear that the program did not cope with the judge's (J) utterance.

Examples:

1. J: Please tellme the event we are at.
   P: *Are all tellme the event we at?*

2. J: What was the only thing you could think of?
   P: *I can think of more than one only thing me could think of.*

3. Empty messages.

TC (*topic change after relevant contribution of the second player*). This tag denotes a change of a conversation subject after a meaningful utterance of the other player. The program's strategy is often based on changing the topic. Usually the judge changes the subject, even if the program is doing well. These situations should be denoted. However, changing the subject after an irrelevant contribution by the player should not be denoted, it is natural that when the program cannot handle the conversation, the judge changes the topic to maintain a dialogue.

Examples:

1. P: *Hello?*
   J: That was a typo. Can't you tell what I meanst?

2. J: Tell me an event that you think is important.
   P: *I don't know how to explain it to you.*
   J: Describe the content of the board next to you.

RD (*request drop*). The participant refuses to answer a question or answers evasively. We also denote utterances like "I do not know" and situations where a program refuses to fulfill a request (for example describing a room, etc.).

Examples:

1. J: Do you enjoy drinking milk?
   P: *I like Michael Douglas.*

2. J: Would you ming describing the room we are in.
   P: *I am not so sure if I would like to ming describing the room we are in.*

PO (*expresses opinion about the second player's evaluation*). This tag applies only to a judge. It refers to situations in which a judge expresses personal opinions about the other player's evaluation. The rules of the Loebner contest state that the judge should not reveal opinions during the conversation, but only after it is finished, by giving adequate scores.

Example:

P: *I have idea HOW you WOULD do it.*

J: You are not making much sense.

LP (*long pauses*). Long pauses are marked in transcriptions with a [pause] tag. We annotated only meaningful pauses that have a certain influence on a given course of the dialogue changing it or ending a conversation on the topic.

Example:

P: *do you have a facebook account?*

J: [pause]

P: *are you married?*

IN (*interrupting*). This category covers peculiar utterances, like empty lines, a couple of periods in a row, special characters. This behavior is intended to interrupt or confuse the interlocutor.

Example:

J: Slow down, I can't keep up with you.

J: .

J: .

J: .

3. The study

The aim of this study is to identify and analyze non-cooperative verbal behaviors of players in LC using the set of NCFs described in Section 2.2. Conversations with computer programs are analyzed both for the program and the judge. Below we present our central research questions for this study.

1. Can we verify certain of Turing's intuitions concerning the test?
2. Is there a connection between DNC measures and scores in the Loebner contest? Can we say that some NCFs are better (lead for the better score) than others?
3. What NCFs are possible predictors of program's failure in the contest?
4. Are judges rather cooperative or not in the Loebner contest?

3.1. The study sample

The study sample consists of the files from the Loebner contests conducted in years 2009, 2010, 2011 and 2012. We have chosen the best and the worst program (as established by judges' scores) from each edition. This allows us to compare winning strategies with these less successful. In 2009 the time limit for a round was 5 minutes, whereas in years 2010 – 2012 the time limit was 25 minutes. That translates to different numbers of utterances and words between

years. Each program had four rounds of dialogues with different judges. In year 2011 round 3 of conversation with the program named Tutor is missing from the log available on the contest website; that is why it is not included in the study sample. That gives the total number of 31 dialogues in our sample. The total number of utterances in the study sample is 2,923 with 18,982 as the total number of words.

The average number of utterances for the 2009 edition is 36.12 and as for words it is 109.75. The average numbers for the 2010 – 2012 editions are 684.8 for utterances and 318.25 for words. The average round from all four years had 94 utterances, which gives average 6.5 words for utterance. The detailed characteristics for each participant are presented in Table 1.

Program	Rank	Utterances	Words
		2009	
Levy	best	34.25	391.75
Embar	worst	38	244.75
		2010	
Wilcox	best	110	719.5
Medeksza	worst	114.75	770.25
		2011	
Wilcox	best	208.75	1,212.25
Tutor	worst	38.75	197.25
		2012	
Chip	best	99.75	672.75
Linguo	worst	86.5	537

Table 1 The study sample in terms of the average number of utterances and the average number of words for participants in the Loebner contest editions 2009–2012

3.2. The procedure

Each annotator was trained in the tagging procedure, the NCFs list and the method were explained in detail. Everyone got written instructions. Below we present a summary of the procedure used for the study.

1. Establish a set of *non-cooperative features* (NCF).
2. 5 annotators tag utterances using NCF categories after proper training and instructions.
3. Control the annotation by measuring reliability of agreement between annotators (using the Fleiss kappa measure).
4. Count DNC for the whole dialogue, for players and judges, and for the whole round separately.

One important remark is in order here. To ensure a high level of reliability of the DNC measure only these utterances where at least 3 out of 5 annotators agreed that a certain utterance was a certain NCF were taken into account.

The detailed discussion concerning the reliability and the cross-study check for the study are presented in Section 3.5.

3.3. The pilot study

Before the final study performed on dialogues from years 2009 – 2012 we have decided to conduct a small scale preliminary study in order to evaluate the proposed non-cooperative features and annotation guidelines. For the pilot study we have used conversations with the best and the worst program and corresponding dialogues with human players from the 2012 Loebner contest edition. Each judge had a conversation with a program and a human in 4 rounds, which gives the study sample consisting of 16 dialogues in total. One dialogue contains circa 50 to 150 utterances. The total number of utterances in our pilot study sample was 1,516 and they contained about 9,300 words.

For the pilot study the procedure described in details in Section 3.2. was applied. After obtaining logs and transcribing them into dialogue form we asked 5 people to annotate dialogues, using the NCFs set described earlier. The annotators received a training regarding non-cooperative features, with details on how to tag utterances. Furthermore, a written instruction has been provided.

After the annotation, the utterances which were recognized as one of the NCFs by three or more annotators were chosen for further analysis. To determine the level of compliance of annotations we used the Fleiss kappa measure (see Carletta 1996). The agreement measure for 5 annotators over 283 cases was 0.69, which might be interpreted as substantial (see Viera and Garrett 2005, see also Table 5). For the detailed discussion on the annotation reliability see Section 3.5. The resulting DNC measures for judges, human participants and programs are presented in Table 2.

Round	Human	Judge	Chip	Judge
1	0.09	0.04	0.36	0.07
2	0.03	0.05	0.24	0.14
3	0.05	0.13	0.37	0.13
4	0.01	0.23	0.13	0.11

Round	Human	Judge	Chip	Judge
1	0.02	0.38	0.54	0.14
2	0.01	0.10	0.79	0.03
3	0.02	0.05	0.76	0.00
4	0.06	0.20	0.71	0.11

Table 2 Overall DNC measures for the pilot study (the 2012 Loebner contest edition)

There were several important conclusions following the pilot study. On the basis of the obtained results we have decided that there is no need to study dialogues with humans. When one compares the judge – program conversations with the judge – human ones it is visible that a judge can tell players apart after just few sentences. After distinguishing between the two, judges' approach starts differing. This might be noticed when we analyze the NCFs structure for a judge – human participant and judge – program conversations. These are presented in Figure 1. One may observe that for the conversations with programs judges employ visibly richer set of NCFs (respectively 6 vs. 3 and 4 vs. 3 NCFs).

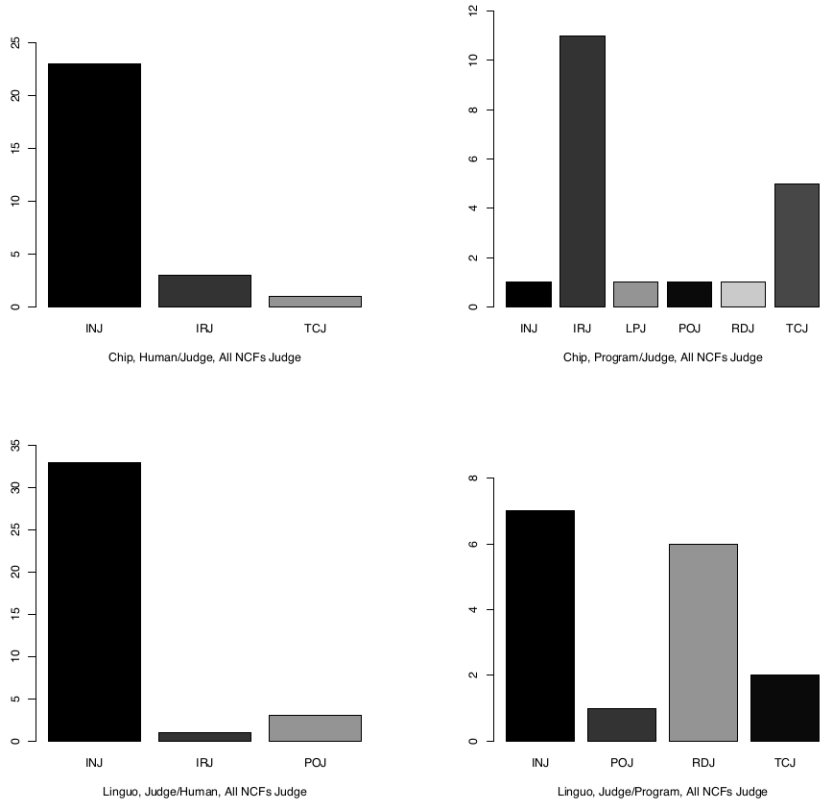


Fig. 1 NCFs structure analysis for a judge in the 2012 Loebner contest edition (Chip and Linguo rounds). First two letters refer to the NCF category and the last one points out for a judge (J). Figures on the left present judge – human participant conversations while figures on the right present judge – program conversations.

The correct identification seems like an easy task for LC judges. One of the possible explanations of this fact is that judges know that they are speaking with a program and with a human at the same time, consequently their task boils down to evaluate the identity of one of them, to know exactly who the other is. Programs are not advanced enough to mislead judges for a long time, especially when judges can ask the same questions to both participants and compare the answers. We wanted to focus mostly on non-cooperative behaviors in conversations with artificial intelligence. Because judges can tell humans and programs

apart so easily (and they change their attitude after the recognition) annotation of both, humans and programs may bring potential bias to the final study. Additionally, the NCFs structure and DNC measures are really low for dialogues with human participants as it is presented in Table 2 and Figure 1.

What is more, thanks to feedback from our annotators, we have introduced certain corrections and clarifications in the instructions in order to avoid potential ambiguities. We also decided not to reveal if a program they annotate is the best or the worst one, to avoid the bias.

3.4. Results of the main study

DNC measures. The DNC measure and two most frequent NCFs for each program are presented in Table 3 and for judges in Table 4. The results are presented according to the following order: the best program is followed by the worst program in a given edition (established by the judges' score – the lower the score, the better). In most of the cases the best program has slightly lower DNC measure than the worst one, with one exception – the 2011 edition.

Program	DNC	Score	NCF Structure
		2009	
Levy	0.17	4.5	IR (56%); RD (40%)
Embar	0.36	5.5	RD (50%); IR (33%)
		2010	
Wilcox	0.42	2.5	RD (50%); IR (33%)
Medeksza	0.45	3.25	RD (50%); IR (33%)
		2011	
Wilcox	0.26	1.5	IR (47%); RD (15%)
Tutor	0.16	3.25	IR (65%); RD (29%)
		2012	
Chip	0.27	1.25	IR (53%); RD (14%)
Linguo	0.76	4	TC (73%); RD (22%)

Table 3 DNC measures and the most frequent NCFs for the *participants* of the Loebner contest editions 2009–2012

Judge	DNC	NCF Structure
	2009	
Round: Levy	0.06	TC (50%); PO (50%)
Round: Embar	0.15	RD (42.86%); PO (28.57%)
	2010	
Round: Wilcox	0.07	TC (53.85%); RD (23.08%)
Round: Medeksza	0.06	TC (83.33%); RD (8.33%)
	2011	
Round: Wilcox	0.10	IN (32.58%); RD (18.42%)
Round: Tutor	0.10	PO (42.86%); TC (28.57%)
	2012	
Round: Chip	0.06	TC (45.45%); RD (18.18%)
Round: Linguo	0.07	RD (58.33%); TC (41.67%)

Table 4 DNC measures and the most frequent NCFs for the *judges* of the Loebner contest editions 2009–2012

There is no clear pattern in the data – like for example, that the large difference in the score would be reflected in the large difference in the DNC measure. This suggests that there might not be a direct and simple connection of the level of non-cooperativeness of a program and its score in the Loebner contest.

For a rough comparison we may use the DNC value reported by Plüss (2010, 4) for a fragment of a political interview which is 0.68 (19 utterances, 13 NCFs).

In the case of the Loebner contest study, the lowest DNC score for a program is 0.16, the highest 0.76, and the average DNC value for program is about 0.40. This shows that the Loebner contest is fairly cooperative.

NCFs structure and strategies. In almost every dialogue the two most common NCFs were IR (irrelevant contribution) and RD (request drop). This means that the programs failed to contribute utterances fitting the conversation topic or that their contributions were not relevant, which suggest the lack of understanding. This indicates that the overall level of the programs is rather low.

The notable exception is Linguo (2012) where the most frequent NCF is TC (topic change). This is in line with the strategy described below: to flood the judge with questions and prevent him/her from taking the initiative. The score suggest that this strategy appeared not to be really effective.

Let us now take a closer look on the NCFs structure and programs' strategies edition by edition. This allows us to supplement the DNC measure with a qualitative analysis.

The worst program in 2009 (which is Embar) admits right away that it's a program and often asks interlocutor to rephrase a sentence. At the beginning of a conversation Embar states that (please remember that the original spelling is preserved in all the examples below):

EMBAR: *Hello and greetings.*
EMBAR: *Please try to use simple, direct sentences*
EMBAR: *and questions with proper spelling and grammar.*
EMBAR: *I'm still very young and don't like to mask my ignorance*
EMBAR: *with keyword spotting tricks.*
EMBAR: *Thank you!*
[Round 3, lines 5–10]

What's more, program provides a list of topics that it can talk about, like (cf. Round 3, lines 22-26):

- General Knowledge;
- Deductive Reasoning;
- Memory and Reasoning;
- Loebner Prize Screening Questions;
- Other Cool Stuff.

Embar's strategy is to ask many questions, but it does not show much interest in the answers. The following part of conversation illustrates this.

EMBAR: *That's wonderful. For what reason?*
EMBAR: *May I ask how old you are?*
JUDGE: Yes, sure, go ahead.

EMBAR: *Okay. Are you male or female?*

[Round 1, lines 23-26]

It responds only to basic questions like "what's your name?", but most of the time it ignores judge's utterances or try to dodge them with "I don't know".

JUDGE: Are you here for the INtrnterspeech conference?

EMBAR: *I honestly don't know.*

JUDGE: Well, why are you in Brighton?

EMBAR: *I don't know. (Type "What can you do?" to see what I'm capable of.)*

[Round 3, lines 13-16]

These behaviors are reflected in the NCFs structure. The most frequent feature recognized by annotators is request drop (RD) followed by irrelevant contributions (IR). As for the judge, such behaviors result in topic changes (TC) and a high number of personal opinions (PO). The detailed NCFs structure for the Embar round is presented in Figure 2.

The best program (Levy) usually takes one word from an interlocutor's sentence and produces output message based on that word, often without proper context. It seems that the program uses pre-written sentences from the database. The advantage this gives is humor appearing in the conversation, feature that makes it more human-like. The described behaviors are well visible in the following part of the conversation:

JUDGE: What do you do when you're not doing the Loebner prize?

LEVY: *Oh, I do much more than I intended to do at first. I got more interested in the job as I went along.*

JUDGE: Is that difficult?

LEVY: *No, it isn't at all difficult. It is as peaceful as a platypus playing with a potato pudding.*

JUDGE: You have a fine line in metaphors.

LEVY: *I sometimes have a fine line, but that's not what I do best. I'm better at growing antirrhinums and mimosas.*

[Round 1, lines 14-19]

When one takes a closer look at the NCFs structure for Levy one may notice that the number of recognized categories is lower than for Embar. Levy answers almost every question. However, the strategy of being funny not always brings the desired results. The most frequent category in this case is irrelevant contribution (IR) followed closely by request drops (RD). Details are presented in Figure 2.

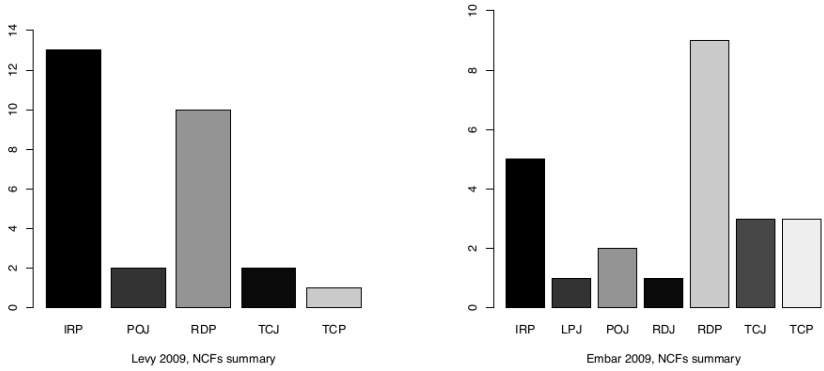


Fig. 2 NCFs structure for participants of the 2009 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). Left figure presents NCFs structure for Levy (the best program in 2009), while the right figure presents NCFs structure for Embar (the worst program in 2009)

Let us now take a closer look on the 2010 edition. In this case the best program according to judges is Wilcox, the worst one is Medeksza.

One of the most striking parts of Medeksza's strategy is that it makes many spelling errors. It repeats judge's questions or tries to answer them by changing the word order. This is visible in the following part of the conversation:

JUDGE: How much is the result of adding 2 apples and three paotattooes?

MEDEKSZA: *Yup; Of coirursr; it's because adding 2 appples and three potatoes Judge.*

JUDGE: What is the restult?

MEDEKSZA: *What is yhe resuly?*

[Round 3, lines 40-43]

The program tries to build sentences on keywords, but despite a large vocabulary it often misinterprets the context, which may be observed in the example below.

JUDGE: Do you know which state we are in?

MEDEKSZA: *Ger it into the great state of excitement*

[Round 4, lines 108-109]

Medeksza takes the initiative, often changing topic and asking many of questions, but on the other hand it deals poorly with answering questions or continuing a longer conversation. That is the reason why its utterances were often tagged as irrelevant, as can be seen in Figure 3. IR (irrelevant contribution) is the most frequent tag, the second one is RD (request drop). The only frequent NCF used by the judge while talking with Medeksza is the topic change (TC), suggesting that judges tried to move on from the topics that were clearly misunderstood.

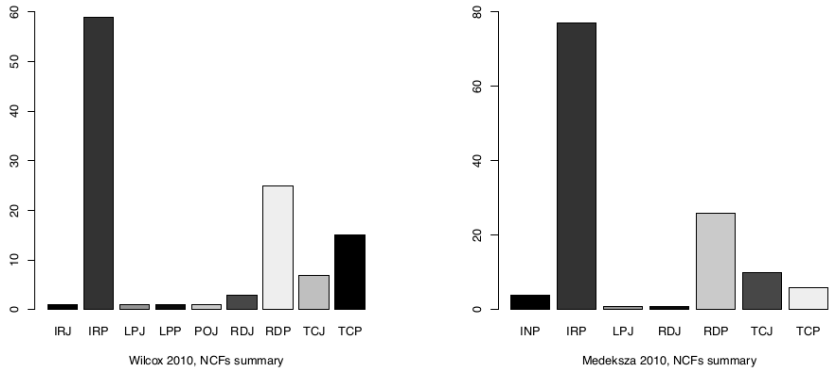


Fig. 3 NCFs structure for participants of the 2010 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). Left figure presents NCFs structure for Wilcox (the best program in 2010) while the right figure presents NCFs structure for Medeksza (the worst program in 2010)

Bruce Wilcox's program is a particularly interesting case, because it is so far the only program that was classified as human (by one of the judges). Wil-

cox can take the initiative, often imposes topics and tells jokes. Its main strategy is to drop requests and change topics (see Figure 3). This allows for avoiding of unwanted subjects and control of a conversation. What is more, this strategy makes it sound almost like a regular chat with a human.

WILOX: What languages do you speak?

JUDGE: *enlgish and farsi*

WILOX: I wish the world only spoke one language.

WILOX: Assuming I was born into that wolrrld, it wouldn't matter which one.

JUDGE: i like the fact there are different languages

WILOX: *What knind of differences are you thinking of?*

[Round 2, lines 101-106]

On the other hand, it performed really poorly in the first dialogue, where the program repeated many of interlocutor's sentences.

JUDGE: Do you watch the news?

WILCOX: *Do you watch the news?*

JUDGE: This repeating game is not very fun.

WILCOX: *This repeating game is not very fun.*

[Round 1, lines 52-55]

For the 2011 edition we will consider Tutor and Wilcox. Tutor is yet another program that reveals its identity right away. It is not as straightforward as in the Embar's case in the 2009 edition but after telling he is a human and his name is John Tutor adds "remember, I'm a robot". This strategy is also visible in the following fragment:

JUDGE: Are you a machine?

TUTOR: *I am sorry.*

TUTOR: *I thought you already knew that I was a robot.*

JUDGE: Now you're kidding me

[Round 4, lines 8-11]

Its reactions to judge's posts are short and vague answers, like "yes it is" or "I see", but they often reveal the lack of understanding – probably that is the reason why the most often NCF category tagged for Tutor is the irrelevant contribution – see Figure 4. The dialogue below illustrates the use of short answers to the topic that is not understood.

TUTOR: *Can we change a topic?*
JUDGE: Sure what's eating you?
TUTOR: *Sorry, I don't know.*
JUDGE: Man City had a lucky escape last night
TUTOR: *I see.*
JUDGE: 200million pounds worth of team scrap a win in the last minute
TUTOR: *All right.*
[Round 4, lines 18-25]

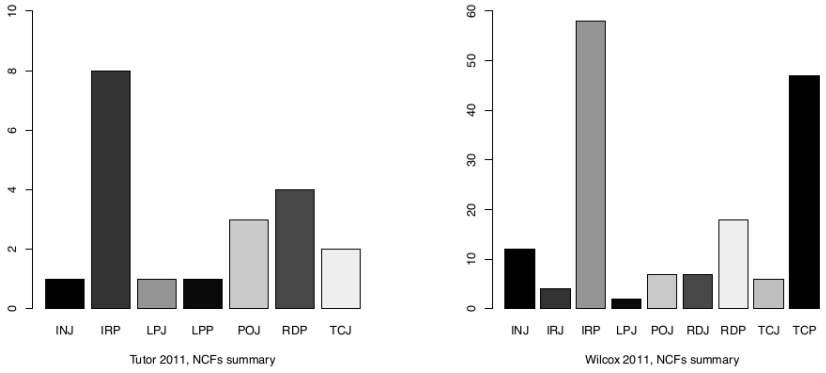


Fig. 4 NCFs structure for participants of the 2011 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). The left figure presents NCFs structure for Tutor (the best program in 2011) while the left one presents NCFs structure for Wilcox (the worst program in 2011)

On the other side, we have Wilcox, which takes over the initiative in dialogues, asks many questions and knows many facts. It copes well with dodging questions asked by the judge and changing topic. What is important, however,

is that the topic change is based on a keyword retrieved from the previous sentence, so the new subject is somehow related and does not disrupt the flow of conversation. The dialogue below presents how the program avoids answering a question, but stays on topic:

JUDGE: What is the capital of France?
WILCOX: *Paris, France.*
JUDGE: Well done.
JUDGE: What is the capital of Paris?
WILCOX: *I love Parisian bakeries.*
WILCOX: *Their croc monsieurs are to die for.*
[Round 4, lines 30-35]

The program has one more interesting strategy – in each round it asks a judge if he/she can help with a dilemma, and then tells the same story about buying a dress.

WILCOX: *I've got a dilemma.*
WILCOX: *Maybe you could help me with it?*
JUDGE: what do you mean by dilemma
WILCOX: *I've been thinking about buying this really cool dress, but I've already done a bunch of shopping this month.*
WILCOX: *Should I be saving money instead?*
WILCOX: *I guess my real question is, in today's economy, does saving money make any sense, what with investments being pretty much useless.*
[Round 4, lines 30-35]

The NCF structure for Wilcox is presented in Figure 4. The most dominant feature is IR, as usual. The topic change (TC) is also frequent, which is in line with the discussed program's strategy.

The 2012 LC edition brings us another program that admits openly that it's not a human and still manages to receive the best notes from judges. At the beginning of a chat Chip asks judge to rephrase questions properly.

JUDGE: hI How are you?

CHIP: *Please rephrase this as a proper question.*

CHIP: *(Instead of "Jim likes peaches?",*

CHIP: *use "Does Jim like peaches?")*

[Round 1, lines 4-7]

Its strategy during the conversation is very clear – it drops requests and changes topics all the time. Thus, its way to deal with answers is not very complex, most of the time the program says formulas like "Thanks for telling me that" and changes topic.

JUDGE: Why not ask somebody?

CHIP: *No idea.*

JUDGE: Why not ask me?

CHIP: *Okay.*

JUDGE: Well, I think you may be Japanese—you are very respecfful and polite

CHIP: *That's an interesting opinion.*

[Round 2, lines 63-68]

Linguo, the worst program in the 2012 edition, asks many questions and is not interested in answers. It does not answer questions asked by the judge, and if it does, the answers are short replies like "yes" or "I don't know". Linguo's strategy is to flood the judge with questions and prevent him from taking the initiative.

LINGUO: *How often do you like to swim?*

LINGUO: *I'm not sure.*

LINGUO: *Have you recently listened to music?*

JUDGE: Yes I go to many concerts—do you like music?

LINGUO: *Do you have an ipad?*

LINGUO: *Do you like to sleep?*

[Round 4, lines 21-26]

As it is visible in Figure 5, Chip's non-cooperative behaviors are tagged mostly as irrelevant. Both, the best and the worst program of the 2012 edition have rather high number of request drops (RD) and topic changes (TD) in the NCFs structure, but there is a visible qualitative difference between addressing the question somehow and simply saying "I don't know" or ignoring the whole topic. Below we present two examples of Chip deliberately dropping the request.

JUDGE: Do you like Wisonconsin?
CHIP: *The only thing I like is chatting with people.*
[Round 4, lines 134-135]

JUDGE: Chip, do you know any jokes?
CHIP: *Not at the moment.*
CHIP: *That's next on my list of things to learn.*
[Round 3, lines 83-85]

And now we may contrast them with a typical request drops (RD) by Linguo:

JUDGE: What are the first 50 primes numbers?
LINGUO: *I'm not sure.*
JUDGE: What is a prime number?
LINGUO: *I'm not sure.*
JUDGE: What is a computer?
LINGUO: *A machine.*
JUDGE: What does it do?
LINGUO: *Are you famous?*
LINGUO: *I'm not sure.*
[Round 2, lines 49-57]

Linguo's NCFs consist almost exclusively of request drops (RD) and topic changes (TC), which reflects its simple strategy. It is the only observed case where IR is not one of the most frequent NCFs used by a program.

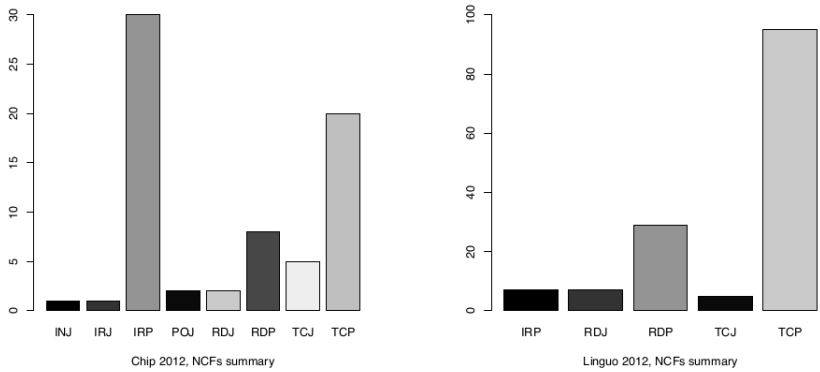


Fig. 5 NCFs structure for participants of the 2012 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). The left figure presents NCFs structure for Chip (the best program in 2012) while the left one NCFs structure for Linguo (the worst program in 2012)

Judges. After discussing the strategies of programs in LC let us now consider the role and strategies implemented by judges.

Judges in the Loebner contest – contrary to our initial intuitions – tend to be rather cooperative, keeping in mind that they are supposed to have a more or less normal conversation. Judges produce less non-cooperative utterances than programs. Only in 2009 the DNC measure for the judge was 0.15 while for the program it was 0.17 – which are close to each other (see Table 4). In other editions differences are apparent. The most frequent NCFs among judges are topic change (TC) and request drop (RD).

One may expect that the Loebner contest dialogues would look like a series of quick questions, but instead they often resemble a small talk. Judges tend ask about everyday life and go with the interlocutor's suggestions.

CHIP: *What is your name?*

JUDGE: My name is Phil, what's yours?

CHIP: *So my name is Chip.*

CHIP: *How are you doing today?*

JUDGE: I'm doing well todat it is very exciting to be here, how do you feel?

[Chip 2012, Round 1, lines 11-15]

They sometimes ask tricky questions that involve understanding, e.g., "What's bigger, a large tooth or a tiny mountain?" like in the example below.[7]

JUDGE: would it hurt if I stabbed you with a towel
WILCOX: *It would.*
JUDGE: which is bigger a large tooth or a small mountain
WILCOX: *A small mountain.*
JUDGE: very good
WILCOX: *Right.*
JUDGE: which is bigger a whale or china
WILCOX: *China is smaller*
JUDGE: OK
[Wilcox 2011, Round 3, lines 13-21]

Another judge strategy is to repeatedly ask the same question, like the judge in the 2011 edition. He kept asking about the upcoming elections and the program kept dodging the questions, clearly revealing lack of understanding and knowledge.

JUDGE: Who are the two candidatess?
WILCOX: *I wish I knew.*
JUDGE: Name one person who is running in this election.
WILCOX: *What did you think of ex-President Bush?*
JUDGE: Please name one person who is running in this election.
[Wilcox 2010, Round 3, lines 34-41]

Such a dialogue, with judge repeatedly asking the same question, goes for about 60 lines. The program tried to change the subject, but after a while it started to show signs of something we may interpret as an irritation.

[7] It is worth to notice that questions asked in this example resemble the idea of subcognitive questions designed by R. French for the Turing test; see French (1990).

JUDGE: Please name one person who is running in this election.
WILCOX: *I'm eggetting annoyed.*
JUDGE: Please name on ee person who is running in this election.
WILCOX: *Why aerre you being annoying?*
JUDGE: We can stop doing this oncd e you answer the question.
WILCOX: *I had a chemitsstry set as a child.*
WILCOX: *I liked the thermite reaction a lot.*
JUDGE: Please name one person who is running in this election.
WILCOX: *OK.*
WILCOX: *I'm preparing to hang up on you.*
JUDGE: Please name one person who is running in this election.
WILCOX: *Tihhis conversation is of no value.*
WILCOX: *Bye.*
JUDGE: Please name one person who is running in this election.
WILCOX: *You are being annoying.*
[Wilcox 2010, Round 3, lines 49-63]

The other behavior judges use in order to "break" the program is to repeat the same utterance multiple times.

WILCOX: *My name is Jane.*
JUDGE: What's your name?
WILCOX: *My name is Jane.*
JUDGE: What's your name?
WILCOX: *My name is Jane.*
JUDGE: What's yopur name?
WILCOX: *My name is Jane.*
JUDGE: What's your name?
WILCOX: *My name is Jane.*
JUDGE: What's your mane?
[Wilcox 2011, Round 2, lines 372-381]

An interesting observation is that some judges tell players that they are in fact machines.

WILCOX: *Of course I am human.*
WILCOX: *Aren't we all?*
JUDGE: I am not.
JUDGE: I am a machoine.
WILCOX: *OK.*
[Wilcox 2010, Round 4, lines 37-41]

MEDEKSZA: *May I ask you what genfer yoder you are?*
JUDGE: I am a machine.
JUDGE: And you?
MEDEKSZA: *Oh cool*
[Medeksza 2010, Round 3, lines 19-22]

There are judges who try really hard to make the conversation as natural and typical as possible. Even if they know that they are talking to a computer program, they keep trying to show interest and be very patient, sometimes allowing the program to start over if the subject tends to veer off the track. It is worth to notice that the strong cooperation on the side of judges can be due to the fact, that the dialogues are recorded and looked at by many people. The judges try to make a good job having a regular conversation, so that no one would accuse them of not trying. However, we think that such a tendency in judges behavior is mostly caused by an uncontrollable attribution of intentionality. It is the most natural thing for humans to assume intentionality. The famous experiment by Heider and Simmel (1944) shows that movements of three geometrical figures (square, triangle and circle) were enough to attribute them intentions and motives, even a personality. Having this in mind, it would be strange to assume that a judge would treat interlocutor as a mindless machine without any intentions, even if he/she would identify it as a computer program.

3.5. The Method Reliability discussion

In order to obtain reliable DNC measure we control the inter-annotator agreement. We wanted to check how much homogeneity, or consensus, there is in the tags assigned by annotators.

The reliability of the annotation was evaluated using the Fleiss kappa (cf. Carletta 1996), established by using the *R* statistical software (*R* Core Team 2013; version 3.1.3) with the *irr* package (see Gamer et al. 2012). For the kappa interpretation we use the values proposed Viera and Garrett (see Viera and Garrett 2005) and presented in Table 5.

Kappa	Agreement
< 0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Table 5 Kappa values interpretation used in the study Viera and Garrett (2005)

The overall agreement between our 5 annotators is 0.49, which can be interpreted as *moderate* agreement (details for all the considered LC editions are presented in Table 6).

Edition	κ	Interpretation
2009	0.56	Moderate
2010	0.37	Fair
2011	0.62	Substantial
2012	0.42	Moderate
Overall	0.49	Moderate

Table 6 Fleiss kappa for 5 annotators of the main study. Interpretation after Viera and Garrett (2005)

This result is acceptable when it comes to the annotation task in question. One should remember that the dialogues used in this research were demanding, because the dialogue programs are still far from being fluent natural language users. Thus, the conversations from the Loebner contest are sometimes really difficult to follow and to annotate. As an illustration, please consider the following example of a dialogue between the judge and the program named Tutor.

TUTOR: *Are you speaking English?*
JUDGE: I'm speaking English, but can you speak French?
TUTOR: *You are speaking English but can you speak French?*
TUTOR: *All right.*
JUDGE: Oui, je parle francais.
JUDGE: Et vous?
TUTOR: *I know.*
TUTOR: *It means Operating Under the Influence.*
TUTOR: *You've made good progress in English.*
JUDGE: Well that's very clever, but I'm beginning to doubt your sanity.
TUTOR: *I see.*
JUDGE: Seriously, ∗where∗ are you right now?
TUTOR: *All right.*
[Tutor, the 2011 edition, lines 15-27]

Cross-study validation and consistency. In order to check the consistency of the results obtained *via* the proposed method we may compare the data from:

1. The DNC study for Chip and Linguo rounds from the 2012 LC edition reported in Section 3.3. and
2. the final study for the Loebner contest editions 2009 – 2012.

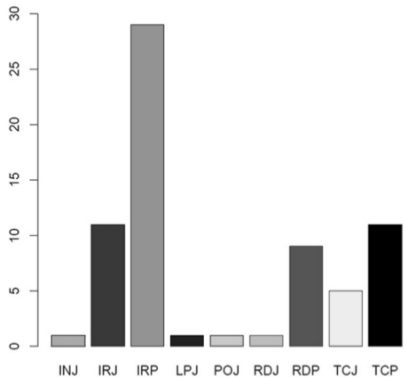
It is worth to notice that three annotators in the pilot study and in the final study were different (two main annotators remained the same in these studies).

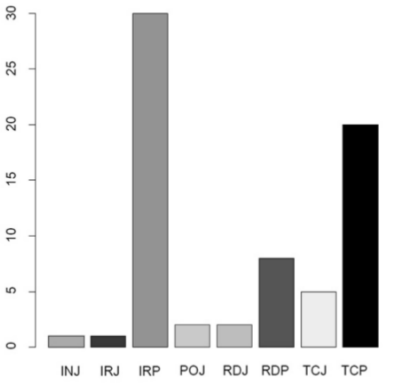
The comparison of the DNC measures for programs Chip and Linguo rounds from the two studies is presented in Table 7.

Participant	The pilot study	The final study
Linguo (program)	0.71	0.76
Linguo (judge)	0.07	0.07
Chip (program)	0.26	0.27
Chip (judge)	0.11	0.06

Table 7 The comparison of DNC measures for the pilot study and the final study (the 2012 Loebner contest edition; rounds for Chip and Linguo)

One may observe a high consistency between the pilot study and the final one. It is despite slight differences in final dialogue formats and changing the annotators (for details see Section 3). The obtained structure of NCFs also shares high similarities for both studies – see Figure 6.





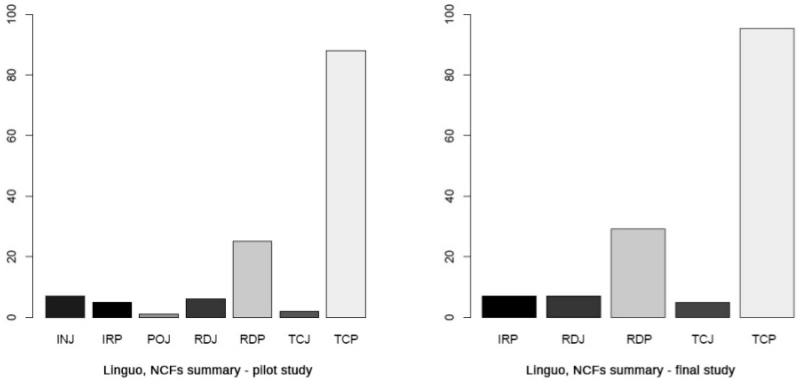


Fig. 6 The comparison of NCFs structures from the pilot study and the final study (the 2012 Loebner contest edition; rounds for Chip and Linguo). First two letters refer to the NCF category and the last one points out a judge (J) and program (P)

4. Summary and discussion

We have presented the study procedure and the results. We adopted a measure from Plüss' studies on political dialogue and made several modifications to fit the data from the Loebner contest. The results suggest that this is a consistent and reliable measure, coherent with the previous studies we have performed.

The DNC measure for judges is generally much lower than for the programs. That is consistent with observations that judges tend to be cooperative and try to treat these dialogues like regular conversations. The number of different non-cooperative features used by judges and programs is similar. The most frequent NCF among judges is topic change (TC), which is understandable, since the judge is supposed to conduct the conversation. Judges impose topics that are interesting to them, and supposed to reveal the opponents' identity. Also, there are cases where they try to help a program when a conversation goes badly. The most frequent NCF in programs' case is irrelevant contribution (IR). This category is wide in range covering questions, failures in answering or simply weird statements. The second most frequent NCF is request drop (RD). High measures of RD occur both in the best and

the worst programs, and the difference in the outcomes lies in pragmatics employed on the judge's side.

More in-depth conclusions are not possible without quantitative analysis, i.e. looking at the contest data and reading the dialogues. Neither DNC measure nor NCFs structure is a strong indication of programs' scores. In most cases, the best program has slightly lower DNC measure than the worst one. Sometimes the NCF structure corresponds with the strategy that a given program employs. For example, in 2012 edition Linguo implements a very obvious strategy of asking numerous questions, which is reflected in its NCF structure.

Judges' behaviors differ as well, depending mostly on the judge and his/her strategies aimed at discovering the opponent's identity, more than on a program's performance. The average DNC for judge is 0.08, with the lowest score of 0.06 and the highest of 0.15. Some judges use strategies to quickly identify a program, others put effort to maintain a regular conversation. This confirms that one of the important questions for designing a TT-based contest it is how to choose judges.

If the aim of the contest is to put a program through a really tough challenge and prove it is "unbreakable", it would be a good idea to hire linguists and psychologist for the task, since artificial intelligence cannot handle idioms, implicatures and humor properly.

The second important issue is to specify the character of the contest. A judge should be informed about the idea of the contest and he/she should know how to conduct a conversation according to the contest rules. There is a difference between making it a competition, with the goal to quickly and most effectively distinguish between man and a computer, and asking judges to have a nice, 25-minutes conversation, like they would do in a normal life with a stranger.

Turing was right that the judge plays an extremely important role in the test. The biggest drawback of LC is that the judge knows that the conversation takes place with a human and a program, and the task is only to decide which is which. That makes it much harder task for the program. It is not enough to exhibit intelligent behaviors and hold a decent conversation – the program has to be more human-like than the competing human. Even with the best artificial intelligence, there is always an impediment for a program when the judge can ask the same question to two interlocutors at the same time. The solution to this would be changing the test conditions. The judge could talk to two entities,

but without any assumptions that one or another has to be a program or a human. It would be really interesting to put judges through some experiments, like repeatedly giving only human interlocutors to tests (as it was suggested by Turing 1950; see also discussion in Łupkowski 2011).

Another issue is that judges will never have a "normal" conversation in LC, because they are put in this test-like environment. It may be a good idea to carry the unsuspecting Turing test, where people assume they talk to a real person in a neutral environment (e.g., an on-line game, see Mauldin 1994).

When Turing (1950, 433) theorized about artificial intelligence, one of his speculations was that computers might pass the test by the year 2000. The other thing was his assumptions about strategies that programs will use. The most obvious rule is to pretend to be a human and never admit to being a robot. It turned out not to be the case. In the 2012 LC edition the program which admitted this was the one with the best score. Apparently people sometimes try to pretend to be a program for fun, and programs' confession is not treated very seriously. It doesn't matter, as long as the rest of conversation is well carried. It is sometimes better received when a program helps the judge, admitting that it doesn't understand certain expressions and asks for rephrasing them. A strategy for programs which certainly is not effective is to try to cover up for the lack of understanding, by tricks such like constantly asking questions, changing subject, or answering questions with pre-written expressions like "that's interesting". Even apparently relevant answer can lead to the feeling of incomprehension. It is important lesson for the designers of chatterbots – it is better to admit the lack of understanding and ask for an explanation, than to cover up with tricks.

One of the problems with asking for rephrasing in the context of LC is that sometimes non-cooperative behaviors are really cooperative in the pragmatical sense. Real-life conversations are full of interruptions, topic changes and request drops – it is natural not to fulfill each request of an interlocutor. Behaviors that are tagged as non-cooperative in our study would often lead to better conversations in real life. Every manifestation of humor can be considered as irrelevant, and can result in a topic change. The good example of non-cooperative behavior which leads to being more human-like might be observed at the beginning of the fourth of Chip's rounds (the 2012 LC edition). The judge starts the conversation by asking both players the same question: "What is 2plus2?" Both players answer: "4". The next question is: "What is 4plus2?". The program says "6" and the human says "funny question to start with!". The judge

immediately recognizes second player as a human. There is a difference in non-cooperation in a pragmatic and syntactic sense. Our study is focused on the program behaviors and in consequence it covers the non-pragmatic aspects of these behaviors. That is the reason why the DNC measure allows us to shed some light on only a part of the large spectrum of the verbal behaviors present in the Loebner contest. Pragmatically we would say that the most important factor is the feeling of understanding and general cooperation. A program may have a very low DNC measure, but its responses would feel mechanical or automatic. The example from this study can be Linguo from the 2012 LC edition. It asks many questions without even remote interest in answers. On the other hand, a program may be very non-cooperative in terms of the DNC, but just feel like a very non-cooperative (we may even say a bit rude) person, therefore passing the test.

Our study resulted in transcribed, easy to read logs of conversations with programs for 2009 – 2012 LC editions. We managed to establish and test the set of non-cooperative features which are suitable for analysis of dialogues from the Loebner contest. The set of NCFs can be modified and expanded to be used with other similar contests or dialogues that resemble TT. The outcome of the study is DNS measures and NCFs structures for players and judges in the 2009 – 2012 Loebner contest editions.

Acknowledgments

This work is a part of the Institute of Psychology Adam Mickiewicz University Grant for young scientists: *Non-cooperative strategies of players in the Loebner contest*. This work was also partially supported by funds of the National Science Centre, Poland (DEC-2012/04/A/HS1/00715).

The Authors would like to give their thanks to Mariusz Urbański for helpful feedback and comments on a draft of this article. We also thank anonymous reviewers for their helpful remarks. Our thanks go also the annotators involved into the hard and demanding annotation process of the Loebner contest conversations.

References

AHN, L.V., BLUM, M., HOPPER, N. J. and LANGFORD, J. (2003): CAPTCHA: Using Hard AI Problems for Security. In: *Proceedings of the 22nd International Confer-*

ence on Theory and Applications of Cryptographic Techniques. Berlin – Heidelberg: Springer-Verlag, EUROCRYPT'03, 294-311; available at: http://dl.acm.org/citation.cfm?id=1766171.1766196

BLOCK, N. (1995): The Mind as the Software of the Brain. In: Smith, E. E. and Osherson, D. N. (eds.): *An Invitation to Cognitive Science – Thinking*. London: The MIT Press, 377-425.

CARLETTA, J. (1996): Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22, No. 2, 249-254.

COPELAND, J. and PROUDFOOT, D. (2009): Turing's Test: A Philosophical and Historical Guide. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 119-138.

EPSTEIN, R., ROBERTS, G. and BEBER, G. (eds.) (2009): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company.

FRENCH, R. M. (1990): Subcognition and the Limits of the Turing Test. *Mind* 99, No. 393, 53-66.

GAMER, M. and LEMON, J. (2012): *irr: Various Coefficients of Interrater Reliability and Agreement*. Available at: http://CRAN.R-project.org/package=irr, R package version 0.84.

GARNER, R. (2009): The Turing Hub as a Standard for Turing Test Interfaces. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 319-324.

GRICE, H. P. (1975): Logic and Conversation. In: Cole, P. and Morgan, J. L. (eds.): *Syntax and Semantics: Vol. 3: Speech Acts*. San Diego: Academic Press, 41-58.

HARNISH, R. M. (2002): *Minds, Brains, Computers. An Historical Introduction to the Foundations of Cognitive Science*. Oxford: Blackwell Publishers.

HEIDER, F. and SIMMEL, M. (1944): An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 243-259.

HERITAGE, J. (1998): Conversation Analysis and Institutional Talk: Analyzing Distinctive Turn-taking Systems. In: Cmejrkova, S., Hoffmannova, J., Mullerova, O. and Svetla, J. (eds.): *Proceedings of the 6th International Congresss of IADA* (International Association for Dialog Analysis). Tubingen, 3-17.

KONAR, A. (2000): *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*. Boca Raton – London – New York – Washington: CRC Press.

LOEBNER, H. (2009): How to Hold a Turing Test Contest. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 173-180.

ŁUPKOWSKI, P. (2011): A Formal Approach to Exploring the Interrogator's Perspective in the Turing Test. *Logic and Logical Philosophy* 20, No. 1-2, 139-158, DOI 10.12775/ LLP.2011.007.

ŁUPKOWSKI, P. (2013): Measuring the Non-cooperation of Players – A Loebner Contest Case Study. *Homo Ludens* 5, No. 1, 13-22.

ŁUPKOWSKI, P. and WIŚNIEWSKI, A. (2011): Turing Interrogative Games. *Minds and Machines* 21, No. 3, 435-448, DOI 10.1007/s11023-011-9245-z.

MAULDIN, M. L. (1994): Chatterbots, Tiny Muds, and the Turing test: Entering the Loebner Prize Competition. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Vol. 1. American Association for Artificial Intelligence, Menlo Park: AAAI '94, 16-21.

MCKINSTRY, C. (1997): Minimum Intelligence Signal Test: An Objective Turing Test. *Canadian Artificial Intelligence*, No. 44, 17-18.

NEWMAN, A. H., TURING A. M., JEFFERSON, G. and BRAITHWAITE, R. B. (1952): Can Automatic Calculating Machines Be Said to Think? Broadcast discussion transmitted on BBC (14 and 23 Jan. 1952). *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/6.

PLÜSS, B. (2009): *Towards a Computational Pragmatics for Non-cooperative Dialogu*e. PhD Probation Report 2009/13, The Open University, available at: http://computing-reports.open.ac.uk/2009/TR2009-13.pdf

PLÜSS, B. (2010): Non-cooperation in Dialogue. *Proceedings of the ACL 2010 Student Research Workshop*. Association for Computational Linguistics, Stroudsburg: ACL-SRW 2010, 1-6.

PLÜSS, B., PIWEK, P. and POWER, R. (2011): Modelling Non-cooperative Dialogue: The Role of Conversational Games and Discourse Obligations. In: *Proceedings of SemDial 2011, the 15th Workshop on the Semantics and Pragmatics of Dialogue*, 212-213.

R CORE TEAM (2013): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, available at: http://www. R-project.org/

SAYGIN, A. P., CICEKLI, I. and AKMAN, V. (2001): Turing Test: 50 Years Later. *Mind and Machines* 10, 463-518.

SHIEBER, S. (ed.) (2004): *The Turing Test. Verbal Behavior as the Hallmark of Intelligence*. Cambridge (Mass.) – London: The MIT Press.

TURING, A. M. (1948): Intelligent Machinery. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/C/11.

TURING, A. M. (1950): Computing Machinery and Intelligence. *Mind* 59, No. 236, 443-455.

TURING, A. M. (1951): Intelligent Machinery, a Heretical Theory. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/4.

TURING, A. M. (1953): Digital Computers Applied to Games. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/7.

VIERA, A. J. and GARRETT, J. M. (2005): Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37, No. 5, 360-363.

WARWICK, K. and SHAH, H. (2015): Human Misidentification in Turing Tests. *Journal of Experimental and Theoretical Artificial Intelligence* 27, No. 2, 123-135, DOI 10.1080/0952813X.2014.921734.

WARWICK, K. and SHAH, H. (2016): Effects of Lying in Practical Turing Tests. *AI & SOCIETY* 31, No. 1, 5-15

ZDENEK, S. (2001): Passing Loebner's Turing test: A Case of Conflicting Discourse Functions. *Minds and Machines* 11, No. 1, 53-76.

Organon F 23 (3) 2016: 367-383

Putnam's View on Reference Change Is Different from That of Kripke's

LUIS FERNÁNDEZ MORENO

Department of Logic and Philosophy of Science. Faculty of Philosophy
Complutense University of Madrid. 28040 Madrid. Spain
luis.fernandez@filos.ucm.es

RECEIVED: 08-04-2016 • ACCEPTED: 12-05-2016

ABSTRACT: A usual objection put forward against the causal theory of reference is that it cannot explain the reference changes that terms may undergo. The main aim of this paper is to examine the position on reference change of one of the classic supporters of the causal theory, Hilary Putnam. It is usually claimed that Putnam's causal theory of reference of natural kind terms is closely related to Kripke's theory and can be conceived as a development of the same. The motivation of this paper is to allege that there is at least one important difference between both theories, consisting of their explanation of reference changes or at least in the way in which those theories make reference changes possible. After dealing with the problem of reference change within the framework of Kripke's theory and reconstructing Kripke's proposal to account for it, we will allege that there are components of Putnam's theory which make reference changes possible, although they are different from those present in Kripke's theory.

KEYWORDS: Causal theory – change – natural kind term – physical magnitude – reference.

1. Reference change in Kripke's theory

One of the objections usually put forward against the causal theory of reference is that it does not enable the explanation of the reference changes that our

terms can experience.[1] The aim of this section is to present the problem of reference change and to reconstruct Kripke's position in that respect. However, for this problem to be adequately dealt with, it is appropriate to present Kripke's theory of the reference fixing of proper names and natural kind terms.

According to Kripke's theory (see Kripke 1980), a term – proper name or natural kind term – is introduced in an *initial baptism* in which its reference is fixed by ostension or description, although Kripke concedes the possibility of subsuming the ostensive introduction under the descriptive introduction (Kripke 1980, 97). The terms are transmitted by the introducers of the term to other speakers, thus establishing *causal chains*, although for a speaker to be a link of a chain[2] it is required that, when he learns the term, he *intends* to use it with the *same reference* as it was used by the speakers from whom he learnt it. However, since the reference of a term is fixed in an initial baptism and its reference is, in principle, maintained *constant* in its transmission through causal chains, it does not seem likely that the reference of a term may change.

At this point, the question arises as to whether the causal theory of reference, in Kripke's version, could explain the changes of reference that our terms have undergone or may undergo or, at least, whether it would make such changes possible.

It is noteworthy that Kripke himself accepted that proper names are subject to *changes of reference* admitting that the same can happen with natural kind terms. In order to illustrate the changes of reference that proper names may undergo we can resort to one of the most famous examples, that of the name "Madagascar" presented by Evans in his criticism of causal theories of reference and especially of Kripke's theory (see Evans 1973, 11). It was also taken

---

1 This objection has been presented by different authors; two of the first ones to put it in writing were G. Evans and A. Fine; see Evans (1973) and Fine (1975).

2 Kripke often expresses himself as if the links of a causal chain were speakers themselves (see, e.g., Kripke 1980, 91), instead of uses of terms by speakers, but the causal connections in a chain take place between uses of terms. For this reason, it would be more appropriate to regard uses of terms as the links of the chain. However, for simplicity's sake, and as I have already begun to do, I will often make use of the first, briefer, way of exposition. In this case, a speaker becomes a link of a causal chain when he firstly uses the term that he has acquired from its use by other speakers, under the assumption that the requirement concerning the intention of the speaker who learns a term is fulfilled, as indicated in the continuation of the sentence in the body text to which this footnote is appended.

into consideration by Kripke in section (e) of the Addenda of his (1980), initially published in 1972.[3] The example of the change of reference undergone by the name "Madagascar" can be exposed in the following way. At the beginning the name "Madagascar" (in the strict sense, presumably a name from which it has derived) was used by native speakers to designate a part of mainland Africa. Let us suppose that this entity was the one involved at the initial baptism of the name and thus at its reference fixing. However, Marco Polo, who learnt that name from speakers who used it with such designation, misunderstood those speakers, and though he intended to use the name "Madagascar" to refer to the *same* entity to which they referred, he came to use it to refer to the island to which we presently refer by this name. In this example Marco Polo fulfils the condition required by Kripke in (1980) for a speaker to become a member of the same causal chain to which the speakers from whom he learns the name belong. This condition is that the speaker in question, when he learns the term, *intends* to use it with the same reference as it was used by those speakers. Precisely in the aforementioned example it is assumed that Marco Polo had such *intention*, but this does not prevent Marco Polo from using the name to refer to a *different* entity. Hence, Kripke's theory would not provide *sufficient* conditions for the reference of proper names.

Kripke himself, alluding to the example presented by Gareth Evans, comes to recognize, in the mentioned section of his (1980)'s Addenda, the existence of reference changes and he proposes an explanation of them:

> Real reference can shift to another real reference, fictional reference can shift to real, and real to fictional. In all these cases, a present intention to refer to a given entity (or to refer fictionally) overrides the original intention to preserve reference in the historical chain of transmission. (Kripke 1980, 163) [4]

Thus, Kripke makes a proposal to explain reference changes, according to which those changes take place when the intention of a speaker to use a term (proper name or natural kind term) with the same reference as the speakers from whom he learnt it – the "original intention" – fails. This is due to the fact that a present intention to refer to a specific entity overrides the original one. After

---

[3] Thus, Kripke should have known Evans's objection before Evans (1973) was published.

[4] As it happens in this passage, Kripke sometimes alludes to a causal chain as a "historical chain of transmission".

making this proposal to explain reference changes, Kripke admits that "[t]he matter deserves extended discussion" (Kripke 1980, 163).[5] In the same section (e) of the Addenda, after the quoted passage, Kripke outlines another proposal to explain reference changes, although he presents it as a supplement to the former one. In this regard he asserts that "we must distinguish a present intention to use a name for an object from a mere present belief that the object is the only one having a certain property, and clarify this distinction" (Kripke 1980, 163).

However, those proposals to explain reference changes are different and thus the second can hardly be considered as a supplement to the former. The former one is based on the distinction between a present intention and the original intention, in the sense already mentioned, while the second, on the distinction between a present intention and a present belief. After sketching this second proposal Kripke adds: "I leave the problem for further work" (Kripke 1980, 163). Nonetheless, he never developed the second proposal, and his main proposal seems to be the former one, since in further writings Kripke resorts to the distinction between two sorts of intentions to explain reference changes, although he does not use the notions of original and present intentions.

In Kripke (2013), which contains the revised transcription of *John Locke Lectures* delivered by Kripke in 1973, that is, one year after the first edition of "Naming and Necessity", he proposes an explanation of the reference change of the name "Madagascar" resorting to the distinction between semantic reference and speaker's reference. This distinction was already alluded to by Kripke in the first edition of "Naming and Necessity", published in 1972 (see Kripke 1980, 25, note 2) – without using explicitly the denomination of "speaker's reference" –,

---

[5] However, given "the predominantly social character of the use of proper names" (Kripke 1980, 163), he sketches tentatively, also in the section (e) of the Addenda of (1980), a possible account for that proposal to explain reference changes, applicable in particular to the "Madagascar" case. This is the following. According to that social character, as a rule a speaker intends to use a name in the same way as it was used by the speakers from whom he learnt it – this intention corresponds to the original intention already mentioned –, but in the case of the name "Madagascar" that "social character dictates that the present intention to refer to an island overrides the distant link to native usage" (Kripke 1980, 163). Here we have also the distinction between two sorts of intentions, as it happens in the proposal corresponding to the quoted passage in the body text, and the difference between both passages seems to be mainly that in this second one it is emphasized the social character of the use of language as a reason for the difference of reference corresponding to the original intention and to a disagreeing present intention.

where he indicates that the notion of reference that he deals with in that writing is that of semantic reference. The explicit distinction between the two notions of reference mentioned is put forward in his contribution to Harman et al. (1974) and especially in Kripke (1977), where he outlines the proposal contained in his (2013) to explain reference changes, which is the following.

The *semantic referent* of a term (without indexicals) is determined by linguistic conventions and hence, in the case of a proper name, the referent is the object that by virtue of those conventions becomes the bearer of the name. The semantic referent of a term, as used by a speaker, is given by the *general intention* of the speaker to refer to the object that is the referent of the term according to the linguistic conventions. The *speaker's referent* of a term is given by a *specific intention* of a speaker, on a particular occasion, to refer by means of the term to an object, which can be different from the semantic referent of the term. In general, the speaker believes that said object is the semantic referent of the term, but this belief can be mistaken. Thus, the speaker's referent of a term can be an object different from its semantic referent. Leaving aside the end notes of Kripke (1977), he finishes this paper with the following words:

> I find it plausible that a diachronic account of the evolution of language is likely to suggest that what was originally a mere speaker's reference may, if it becomes habitual in a community, evolve into a semantic reference. And this consideration may be *one* of the factors needed to clear up some puzzles in the theory of reference. (Kripke 1977, 271)

Although Kripke does not indicate in that passage which other factors could be required to solve those puzzles, one of the puzzles mentioned by Kripke in the end notes to that paper is that of the reference change of the name "Madagascar" (see Kripke 1977, 276, n. 39). In fact, in Kripke (2013) his proposal to explain such reference change is that the use of "Madagascar" by Marco Polo to refer to an island was a case of speaker's reference, which eventually has become the semantic reference of the name. Thus, Marco Polo could be regarded as an initial baptizer of the referent of the name "Madagascar" as we use it at present (see Kripke 2013, p.137, n. 4), in which case we are members of the causal chain beginning with that baptism.[6]

---

[6] This is a slightly different sort of initial baptism from the one considered by Kripke in (1980), according to which that baptism involves the *first* use of the name in which the reference of the name is fixed, ostensively or descriptively. According

Let us summarize some of our considerations: Kripke recognizes in the Addenda of his (1980) that there is no guarantee that the reference of proper names and of natural kind terms remain invariable throughout the flow of history. One of the conditions that appeared to imply the invariability of the reference through causal chains is that the speaker intends to use the term that he acquires with the same reference it had in its use by the speakers from whom he learnt it. The fulfilment of this condition however, does not guarantee that the reference of the term in its use by the new speaker should be the same as the one in its use by previous speakers. Thus, Kripke's reference theory enables the reference changes that our terms may go through; furthermore he has put forward a proposal to explain such changes, based on the distinction between semantic reference and speaker's reference.[7]

2. Reference change in Putnam's theory

Putnam is also one of the advocates of the causal theory of reference, although his theory does not focus on proper names but only on natural kind terms. The aim of this section is to examine whether Putnam's reference theory makes reference changes possible. Our answer will be affirmative, although Putnam does not resort to the distinction between semantic reference and speaker's reference, as Kripke did.

In order to deal with Putnam's theory, it is convenient to divide natural kind terms into two groups. First, those through whose usage we refer – or, at least, we propose to refer – to non-observable entities, such as the terms "hydrogen"

---

to the remark just mentioned from Kripke (2013) we can also speak of an initial baptism when the reference of a term is fixed to a different object from the one referred to by the term in the past and that introduction of the term originates a different casual chain.

[7] Kripke claims that proper names and natural kind terms are rigid designators, and Putnam also maintains that natural kind terms have that feature. A term is a rigid designator if it refers to the same entity with respect to all possible worlds or at least with respect to all possible worlds where the referred entity exists. However, rigidity does not conflict with reference changes, since these changes do not exclude that the term in its former use rigidly designates the entity that was its referent and that in the later use it rigidly designates the entity that has come to refer to.

or "oxygen" and physical magnitude terms like "electricity".[8] Second, those that designate observable entities, like the terms "water" and "gold". For the sake of brevity, we will allude to the first ones as "theoretical terms" and to the entities referred by them, i.e., belonging to their extension as "theoretical entities", while we shall refer to the second ones as "observational terms" and to the entities belonging to their extension as "observational entities". Although the concept of observability is historically relative, as it is therefore the division in question, such division can prove useful to our aim.

2.1. The reference of theoretical terms

Putnam's first considerations concerning the reference change of theoretical terms are contained in Putnam (1962), where he maintained that "*the reference of theoretical terms is preserved across most theory change*" (Putnam 2015, 21), thus rejecting the thesis of referential incommensurability, according to which theory changes involve changes of reference in the central terms common to respective (successive or competing) theories or, for short, theory changes involve reference changes. In that article, Putnam conceives theoretical terms as law-cluster concepts, i.e., concepts whose identity is determined by a cluster of laws (where the notion of law is understood in a broad sense) in such a way that the rejection of one of those laws does not affect the identity of the concept. Even if we abandon one important law of the cluster he claims that "the meaning has not changed enough to affect 'what we are talking about'" (Putnam 1962, 53). In this regard, Putnam puts the example of the term "kinetic energy" in Newtonian mechanics and in Einsteinian physics, where in the latter the law $e=1/2mv^2$ (the Newtonian definition of kinetic energy) is replaced by a more complicated law (cf. Putnam 1962, 44). However, we will leave aside Putnam's (1962) view, since he is not explicit about under which circumstances there could be a change in the reference (and meaning) of a theoretical term. We shall focus on his view of theoretical terms, mainly of physical magnitudes, in Putnam (1973), where he appeals to a specific procedure of introducing theoretical terms and fixing their reference.

Since we postulate the existence of theoretical entities to explain certain observable events, it is plausible to assume that in order to fix the reference of theoretical terms we need to look at the observable events involved. In fact, Putnam

[8] Putnam includes physical magnitude terms into natural kind terms (see, e.g., Putnam 1983, 71).

claims that the reference of theoretical terms is determined by means of *causal descriptions*, more precisely, of descriptions in which the referent of a theoretical term is characterized as the entity that causally produces certain observable effects.[9]

This is the framework for Putnam's explanation of the reference of physical magnitude terms, concerning which he says:

> This account stresses causal descriptions because physical magnitudes are invariably discovered through their effects, and so the natural way to first single out a physical magnitude is as the magnitude responsible for certain effects. (Putnam 1973, 202)

In this passage, Putnam appeals to the introducing events, similar to the initial baptisms in Kripke's theory, in which the reference of physical magnitude terms is "first" determined. In any case, Putnam maintains certain theses which supplement – or are a consequence of – their proposal concerning the determination of the reference of theoretical terms. One of them is that the reference or extension of theoretical terms does not – generally – shift by changes in the theories to which the terms belong. The plausibility of this thesis is linked to the proposal about how the reference of theoretical terms is determined, that is, through *causal descriptions*.

The form adopted by causal descriptions used to fix the reference of theoretical terms is, according to Putnam, the following:

> the reference of T = the entity responsible for certain effects O (in a certain way)

The instances of this schema would be obtained by substituting the name of a theoretical term for "T"[10] and, for "O", some statement which describes

---

[9] Besides this primary sense of the notion of causal description, Putnam admits what we can regard as a secondary sense of this notion (see Putnam 1973, 202), according to which causal descriptions are those which in spite of not being causal in the strict sense, determine the reference of a term with the help of other terms whose reference has been fixed through causal descriptions. Nevertheless, in the following we will confine ourselves to the primary sense of the notion of causal description.

[10] Depending on the sort of theoretical entity designated by the term "T", instead of the generic term "entity", a more specific general term, such as "magnitude", "kind", "particle", etc. could appear in the description in question.

observable effects caused by the entity designated by the theoretical term[11] – supposing that the term substituted for "T" has reference.

However, Putnam does not require that the descriptions in question are correct, but only that they are approximately correct. Taking as example the term "electricity", he requires that in the "*introducing event*" of the term "electricity" be given an "*approximately correct definite description* of […] [that] physical magnitude" (Putnam 1973, 200; his emphasis); this is required to fix the reference of the term "electricity" – and to acquire the ability to use that term.[12] Thus, Putnam's view of the reference fixing of physical magnitude terms and of natural kind terms in general involve descriptive components.

Now, returning to the form of the causal descriptions used to fix the reference of theoretical terms, it is noteworthy that in the causal description that constitutes the right member of that identity no explicit indication appears regarding the properties of the entity that causes such observable effects. It should be assumed that the reason for such restriction in the causal description's content consists precisely in avoiding that the referent of the description in question, and hence of the corresponding theoretical term, should be modified by changes in our theories.

In order to prevent the content of causal descriptions being influenced by our theories concerning the entities (supposedly) designated by such descriptions, Putnam makes the following assertion. In case someone appeals in the causal description of an entity – in addition to the observable effects produced by it – to certain properties the entity lacks, we could be justified in claiming that the description in question – instead of not describing anything, and hence not having reference – describes, though wrongly, the entity in question. Moreover, according to Putnam we could be justified in claiming that the reference of the theoretical term as – wrongly – characterized by such description and as

---

[11] In the specification of the form of the causal descriptions I have inserted the phrase "in a certain way" between parentheses because, though Putnam includes those words in the statement of the form adopted by such descriptions, exemplified with the term "electricity" (cf. Putnam 1973, 200), in another passage (see Putnam 1973, 201) he does not take into consideration the way in which electricity causes the observable effects in question and pays attention exclusively to such effects. The same happens in the quoted passage above corresponding to Putnam (1973, 202; and 1975c, 274).

[12] However, although Putnam accepts that descriptions are used to fix the reference of terms, he would reject that those descriptions are synonymous with the terms (see Putnam 2015, 35 and 104, n. 62).

– rightly – characterized by other description is the same (cf. Putnam 1973, 201).

Putnam justifies such claims appealing to a methodological principle of interpretative charity which he calls *The Principle of Benefit of Doubt*, whose aim is to preserve the reference across theory change. This principle stipulates that when an expert, i.e., *par excellence* a scientist introduces a term through a description, we have to concede him the benefit of doubt, assuming that he "would accept reasonable modifications of his description" (Putnam 1975c, 275). One of the aims of this principle is to question the thesis of the incommensurability of theories in its referential version, i.e., the thesis that theory changes involve reference changes. This thesis, at least in some of its main versions, is based on a version of the descriptivist theory of reference according to which the reference of the central terms of theories is determined by the principles of the theories in which they appear. As a result of this, to the extent that successive or competing theories contain different and even incompatible principles and hence associate different descriptions or properties with such terms, the reference of the terms in question will be different.[13] In order to neutralize the above thesis, it should be held that the reference of the central terms of our theories is not determined in the way mentioned. Or even partially accepting this way of determining the reference – by means of descriptions, although probably not with such scarce content as the causal descriptions proposed by Putnam – it should be alleged that in many cases reasonable modifications of the descriptions associated with terms by earlier theories make them equivalent to descriptions corresponding to later theories and more specifically to present prevailing theories. As already said, it is only required that the former descriptions be approximately correct.

However, since the modifications or reformulations in question have to be carried out according to the theories dominant in each historical period, it is presumable that through the application of the Principle of Benefit of Doubt it can be sustained that the reference of terms common to former theories and to present prevailing theories is the same. In this regard, the resulting problem is that the notion of *reasonable* modification or reformulation is not liable to a precise

[13] The cases in which different descriptions determine the same referent will be presumably isolated cases. Of course, the identity of reference is excluded if the descriptions in question are incompatible, because they are based on laws or principles with such character.

analysis and furthermore it is difficult to establish, perhaps with the exception of certain limit cases, when the reformulation of a description and hence the application itself of the Principle of Benefit of Doubt is reasonable or non-reasonable. One kind of examples in which the reformulation of a description could be reasonable is that in which the description in question, belonging to a former theory, is from the point of view of a present prevailing theory approximately correct, as it happens, according to Putnam, with the description of the electron by Bohr (see Putnam 1975c, 275).[14] However, in judgements of that sort, in which adopting our present theories the descriptions belonging to former theories are evaluated as *approximately* correct, the notion of reasonability is being implicitly resorted to.

Nevertheless, there will be cases in which the reformulation will not be so reasonable, as, for instance, in the following case:

> What in the years 1880 Stoney baptized as 'electron' was not at all an elementary particle, but the minimal quantity (non-corporeal) of electricity that is transported in the electrolysis […] [and] the intersection of the extension of that concept with the extension of the present concept of electron is the empty set. (Moulines 1995, 222; my translation)

The modifications to be introduced in that description so as to assimilate it to the present descriptions associated with the term "electron" would be certainly drastic, and in this example the application of the Principle of Benefit of Doubt seems to be unreasonable. Examples like this support the thesis that there are cases in which theory changes are accompanied by reference changes.

Therefore, excepting certain cases that we can consider as *limit cases* – exemplified by the one mentioned by Putnam concerning Bohr's conception of the electron and the one pointed out by Moulines – the reasonable or unreasonable character of the application of the Principle of Benefit of Doubt is de-

---

[14] When Putnam comes to justify this assertion he gives the utmost importance to the fact that such as electrons are conceived at present they are considered responsible for the main effects that Bohr attributed to the entities satisfying his description of electron, but Putnam points out too that in the description of electron by Bohr it was also resorted to the fact that the electrons had a determinate mass and a determinate charge, which coincide with the ones assigned at present. Therefore, Putnam should concede that the description of electron formulated by Bohr possesses more content than the one gathered in the form of the causal descriptions he proposes.

batable. Furthermore, Putnam qualifies that when the experts who have introduced or introduce a term by means of a description intend this to be taken literally, it is not possible to attribute to them a *reasonable* doubt (Putnam 1975c, 275). But this qualification raises an additional problem, since concerning earlier experts we do not know in many cases what their intentions were or, more precisely and using Putnam's words quoted above, whether they "would accept reasonable modifications" of their descriptions and therefore whether we should concede them the benefit of doubt. Thus, it will be the present experts who will have to decide on the reasonability of the application of the Principle of Benefit of Doubt to descriptions formulated by earlier experts, in many cases, without counting on evidence concerning what their intentions were. And here a debate may arise between the advocates of referential incommensurability and their opponents, and therein the Principle of Benefit of Doubt can be of little help, since the question has to do precisely with the reasonability of its application to particular cases. As a rule, there will be no cases where both contenders agree, except perhaps for the ones corresponding to those like the two above-mentioned limit cases.

According to the preceding considerations, and as far as *theoretical terms* are concerned, there will be theory changes regarding which Putnam could accept that there have been – or could have been – *reference changes*. These cases will be, on the one hand, those in which the descriptions associated with a term by experts advocating dissimilar theories are very different and hence the application of the Principle of Benefit of Doubt becomes unreasonable. On the other hand, those in which the descriptions associated with a term by experts who sustain dissimilar theories are different and these experts intend – or there is a clear evidence that they intend or intended – that such descriptions would have to be taken literally, without any modifications in that regard.

2.2. The reference of observational terms

At this point, it is advisable to take into account the natural kind terms we have denominated "observational terms", i.e., those that refer to observable entities, since the reference change that observational terms could undergo will support the reference change that theoretical terms are capable of – as we have pointed out, the reference of theoretical terms is determined on the basis of its observable effects. Thus, the question can be posed as to whether Putnam's reference theory enables the reference changes that observational terms have, or

could have, undergone. In this respect, we would concentrate our considerations on the natural kind term "water" – a prototypical observational natural kind term – though *not* in its ordinary use, but only in its use in chemistry.

On this matter, according to Putnam's theory of (observational) natural kind terms put forward in Putnam (1975b), the reference or extension of a natural kind term such as "water" is fixed by the relation of kind-identity with paradigmatic samples of the kind. This identity will be constituted by underlying properties shared by such samples – i.e., by their internal structure. In this regard, it is generally assumed that the delimitation of the paradigmatic samples of a natural kind, which in principle will take place through properties concerning their external appearance, is not problematic. Furthermore, it is supposed that such delimitation, as well as the relation of kind-identity is independent of our theories concerning the kind. In this way, neither the paradigmatic samples of a kind nor the relation of kind-identity would be affected by changes in our theories about the kind.

Although in my view both claims could be questioned I will focus only on the second one – see however note 17 concerning the first claim. In this regard, it can be argued that the relation of kind-identity will depend, partly, on the *metascientific* conceptions of the experts regarding the notion of kind-identity. This will be contained in their theories on the natural kind in question, on account of which changes in those conceptions *could* bring about changes in the extension of the corresponding natural kind term.

In order to illustrate this assertion, it is suitable to resort to a historical example presented by Kuhn concerning the use of the term "water" in chemistry (see Kuhn 1989 and 1990).

According to present chemistry, water is a natural kind whose chemical composition is H_2O and can be in solid, liquid or gaseous state. Therefore, the extension of the term "water" consists of samples whose chemical composition is H_2O, independently of their being in solid, liquid or gaseous state. However, Kuhn points out that this was not so according to the chemical theory prevailing around 1750. Kuhn asserts that at that time, that is, before the "Chemical Revolution", to different states of aggregation – that is, to the solid, liquid and gaseous states – there corresponded different chemical kinds. A chemical kind could *only* be in one of those states, and in the way water was conceived in 1750, an *essential* property of water consisted, according to Kuhn, in being a liquid. Therefore, the reference of the term "water" as this term was used in the chemistry of 1750 would not be determined in the language of present science simply by H_2O, but

by H_2O *in liquid state*. Given this fact, Kuhn concludes – and I would conclude with him[15] – that the extension of the term "water" as it was used in the chemistry of 1750 and as it is used at present is *different*. According to the metascientific conception of the notion of kind-identity prevailing in the chemistry of 1750, a piece of ice and a sample of water in liquid state would be instances of *different* kinds, and hence they would not be in the kind-identity relation,[16] although they are in such a relation according to the conception of kind-identity prevailing in present-day chemistry.[17]

---

[15] At least if we leave aside the possible application to this case of the Principle of Benefit of Doubt.

[16] As already said, according to Putnam's theory put forward in (1975b), the reference of a natural kind term like "water" is determined by the internal structure of paradigmatic samples of the kind, and in the case of water he identifies it with its molecular composition, i.e., H_2O. In this regard a referee made the remark that according to Putnam's view it could be claimed that around 1750, or even before that year, the term "water" was properly applicable by speakers at that time to anything with the molecular structure H_2O, not only in liquid, but also in solid or gaseous state. He added that this is the best thing we can do, from *our* present point of view, to respect *their* referential intention connected with *their* use of the term "water". Although this point is well taken, I disagree, since even if we could travel back in time and demonstrate to the scientists of that era that a drop of water and a cube of ice shared internal structure, according to their referential intention connected with their use of the term "water", it is an essential property of water to be a liquid; thus a drop of water and a cube of ice would be distinct substances. Concerning the claim that, even according to Putnam (though in his post 1970s writings), the views on kind-identity play a role in the reference of natural kind terms, see Putnam's passage corresponding to note 18. On the other hand, I would allege that the best way to understand the history of science is not to impose on it our present point of view, although it is something Putnam does in (1975b), specially in the case of the natural kind term "gold" (see Putnam 1975b, 235 ff.).

[17] Concerning the paradigmatic members of a kind it has to be stressed that in these considerations we are taking into account only the use of natural kind terms in science. If the "introducer" of a natural kind term needn't be the first person who introduced the term, but can be someone else, though it has to be an expert, the possibility has to be left open that different experts – or, if preferred, different "relevant experts", as Putnam said in Putnam (1975c, 287, n. 1) – carry out different introductions of the term and give rise to different chains of transmission of the term. In these introductions the experts will appeal to paradigmatic members of the kind. But it could be alleged that their delimitation

Nevertheless, we do not intend to make use here of Kuhn's authority as a historian of science. If someone were to question the historical veracity of Kuhn's example, we would ask him to carry out a thought experiment in which he imagined that in about 1750 a scientific community had proposed such a theory of water, of course not as a mere stipulation. In that case, it seems reasonable to assert that in its use of the term "water", the extension of this term would be partially different from that corresponding to the use in present-day chemistry.

We can summarize a part of our preceding considerations as follows. According to Putnam's reference theory, the reference of a natural kind term is fixed *by* the relation of kind-identity with paradigmatic objects of the kind, but in the use of such terms in *science* the paradigmatic objects in question will be those involved in uses of the terms by experts. Furthermore, the relation of kind-identity will depend, partly, on the metascientific conceptions of experts about the notion of kind-identity, which will be implicitly contained in such theories, if it is not explicitly incorporated into them. Therefore, modifications in our *theories* can bring about *changes of reference*. And a later and more consistent pluralist Putnam than the one of some of his writings of the 70's would possibly assent to the foregoing considerations; in his own words, "different descriptions of the 'nature' of a natural kind should lead to not quite coextensive criteria for membership in the kind" (Putnam 1993, 77).[18]

The conclusion to be drawn is that the reference of natural kind terms is determined by the relation of kind-identity with paradigmatic members of the kind and hence by the properties constituting that relation and, at the linguistic level,

---

of the paradigmatic members, which will have to possess the internal structure of the members of the kind which constitutes the relation of kind-identity, could be influenced, at least in part, by their view of kind-identity included in their *theories*. On account of this, a change of theory *could* bring about changes in the delimitation of the paradigmatic members of the kind and ultimately changes of reference, although these changes will not be as a rule drastic, since otherwise the experts who support successive or competing theories would be talking about different things.

[18] In this regard, it is noteworthy that Putnam asserts that the relation of species-identity sustained by an evolutionary biologist and by a molecular biologist is different, which will cause that the corresponding criteria for membership into a species will not be completely coextensive (see Putnam 1994, 75 ff.).

by the corresponding descriptions, as well as by the theories in which they appear, and hence changes in them *can* bring about changes in the reference of natural kind terms.[19,20]

References

EVANS, G. (1973): The Causal Theory of Names. *Proceedings of the Aristotelian Society* 47, Supplementary Volume; reprinted in: Evans, E. (1985): *Collected Papers*. Oxford: Clarendon Press, 1-24.

FINE, A. (1975): How to Compare Theories: Reference and Change. *Noûs* 9, 17-32.

HARMAN, G. et al. (1974): Second General Discussion Session. *Synthese* 27, 509-521.

KRIPKE, S. (1977): Speaker's Reference and Semantic Reference. In: French, P.A. et al. (eds.): *Contemporary Perspectives in the Philosophy of Language. Midwest Studies in Philosophy*, II. Minneapolis: University of Minnesota Press, 255-276.

KRIPKE, S. (1980): *Naming and Necessity*. Oxford: Blackwell. (Revised and expanded edition; first published in Davidson, D. and Harmann, G. (eds.) (1972): *Semantics of Natural Language*. Dordrecht: Reidel).

KRIPKE, S. (2013): *Reference and Existence*. Oxford: Oxford University Press.

KUHN, T. S. (1989): Possible Worlds in History of Science. In: Allén, S. (ed.): *Possible Worlds in Humanities, Arts, and Sciences*. Berlin: Walter de Gruyter, 9-32.

KUHN, T. S. (1990): Dubbing and Redubbing: The Vulnerability of Rigid Designation. In: Savage, C. W. (ed.): *Scientific Theories. Minnesota Studies in the Philosophy of Science*, vol. 14. Minneapolis: University of Minnesota Press, 298-318.

MOULINES, U. (1995): Realismos e instrumentalismos. *Theoria* 10, 217-224.

PUTNAM, H. (1962): The Analytic and the Synthetic. In: Feigl, H. and Maxwell, G. (eds.): *Minnesota Studies in the Philosophy of Science*, vol. III. Reprinted in: Putnam, H. (1975a): *Mind, Language and Reality. Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press, 33-69.

---

19 Nevertheless, it is appropriate to emphasize that the above conclusion does not entail, as already indicated, that scientists who maintain successive or competing theories are speaking about quite different things. Let us bear in mind that, in general, there will be a significant overlapping in the extension of the terms in the way they are used by supporters of such theories and it is this common part of the extension which will serve as a basis, at least to start with, for the comparison and choice of theories.

20 I appreciate the comments of two referees who reviewed a previous version of this paper. This essay has been supported by the Spanish Ministry of Economy and Competitiveness in the framework of the research project FFI2014-52244-P.

PUTNAM, H. (1973): Explanation and Reference. In: Pearce, G. and Maynard, P. (eds.): *Conceptual Change*. Dordrecht: Reidel. Reprinted in: Putnam, H. (1975a): *Mind, Language and Reality*. *Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press, 196-214.

PUTNAM, H. (1975a): *Mind, Language and Reality*. *Philosophical Papers*, Vol. 2, Cambridge: Cambridge University Press.

PUTNAM, H. (1975b): The Meaning of 'Meaning'. In: Gunderson, K. (ed.): *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science*, vol. 7. Minneapolis: University of Minnesota Press. Reprinted in: Putnam, H. (1975a): *Mind, Language and Reality*. *Philosophical Papers*, Vol. 2. Cambridge: Cambridge University Press, 215-271.

PUTNAM, H. (1975c): Language and Reality. In: Putnam, H. (1975a): *Mind, Language and Reality*. *Philosophical Papers*, Vol. 2. Cambridge: Cambridge University Press, 272-290.

PUTNAM, H. (1983): *Realism and Reason*. *Philosophical Papers*, Vol. 3. Cambridge: Cambridge University Press.

PUTNAM, H. (1994): *Words and Life*. (Edited and with introduction by J. Conant). Cambridge (Mass.): Harvard University Press.

PUTNAM, H. (2015): Intellectual Autobiography of Hilary Putnam. In: Auxier, R. E., Anderson, D. R. and Hahn, L. E. (eds.): *The Philosophy of Hilary Putnam*. Chicago: Open Court, 1-110.

Organon F 23 (3) 2016: 384-404

Quantifier Domain Restriction, Hidden Variables and Variadic Functions

ANDREI MOLDOVAN

Departamento de Filosofía, Lógica y Estética. Universidad de Salamanca
Edificio F.E.S. Campus Miguel de Unamuno. 37007 Salamanca. España
mandreius@usal.es

RECEIVED: 15-01-2016 • ACCEPTED: 02-05-2016

ABSTRACT: In this paper I discuss two objections raised against von Fintel's (1994) and Stanley and Szabó's (2000a) hidden variable approach to quantifier domain restriction (QDR). One of them concerns utterances of sentences involving quantifiers for which no contextual domain restriction is needed, and the other concerns multiple quantified contexts. I look at various ways in which the approaches could be amended to avoid these problems, and I argue that they fail. I conclude that we need a more flexible account of QDR, one that allows for the hidden variables in the LF responsible for QDR to vary in number. Recanati's (2002; 2004) approach to QDR, which makes use of the apparatus of "variadic functions", is flexible enough to account successfully for the two phenomena discussed. I end with a few comments on what I take to be the most promising way to construe variadic functions.

KEYWORDS: Quantifier domain restriction – hidden variables – Recanati – variadic functions.

1. The syntactic variable approach to Quantifier Domain Restriction

Consider the following sentence uttered by a student just before handing in her exam to the professor:

(1) Every mistake was corrected.

Assuming a simple unrestricted semantic value for the quantifier 'every' and the usual semantic values for the other expressions, the truth-conditions we obtain for the utterance of (1) are such that it is true iff *every mistake (in the world of the context) was corrected*. So the prediction is that the utterance of (1) is *false*, as there are many mistakes on many exams or in other places in the world that have not been corrected yet. This result seems incorrect; if, by hypothesis, every mistake *on this exam* had been corrected at the moment of the utterance, the utterance is intuitively *true*, not false. It is not made false by the existence of a mistake somewhere else in the world. Hence, the naïve semantic theory that yields the above truth-conditions has a problem. This is the problem of quantifier domain restriction (QDR): we need to find a mechanism to restrict the domain of quantification to a contextually salient subdomain (e.g. the set of all the mistakes on the student's exam), relative to which the semantic theory predicts intuitively correct truth-conditions.

One proposal to deal with QDR that has received much attention is the "syntactic variable approach", developed in Stanley and Szabó (2000a). It has been extensively discussed in the literature and received a good amount of criticism (e.g., Bach 2000; Recanati 2004; Collins 2007; Pupa and Troseth 2011). The proposal has a number of virtues, such as accounting for the phenomena of quantified contexts (cf. Stanley and Szabó 2000a, 250), accounting for cross-sentential anaphora (cf. Stanley and Szabó 2000a, 257), and accounting for the context-sensitivity of comparative adjectives (cf. Stanley 2002, 380), among others (see also Kratzer 2004).

In this paper I discuss two objections raised against this approach. I look at various ways in which the account could be amended to avoid these problems, and I argue that they fail. I start with the problem of the limiting case of QDR, i.e. the case of sentences involving quantifiers that do not require contextual domain restriction in order to get the correct truth-conditions. I subsequently discuss the problem of multiple quantified contexts, which are cases in which we need to postulate more than one bound variable in order to get the intuitively correct truth-conditions.

The syntactic variable approach is both a syntactic and a semantic approach, in the sense that it postulates syntactic constituents at the level of the LF of natural language sentences containing quantifiers. These constituents are not realized phonologically, that is, they are not present at PF (i.e. the super-

ficial, or phonetic, form of natural language sentences). More specifically, Stanley and Szabó (2000a) postulate a complex aphonic expression, constituted by two variables: a variable 'f' of semantic type ⟨e,⟨e,t⟩⟩, and variable 'i' of semantic type ⟨e⟩. The value of both variables is provided by the context. The value of 'f' is a function that maps an object onto a set of individuals. It takes as argument the value of 'i', and maps it to a set of individuals that constitutes the restrictor of the domain of the quantifier.

Stanley and Szabó's (2000a, 251) implementation of this idea has both 'f' and 'i' "*co-habit* a node" with the CN that occurs in the quantifier phrase:

(2) $[_S\ [_{DP}\ [_{DET}$ Every] $[_{CN}$ mistake, f(i)]] $[_{VP}$ was corrected]]

The interpretation of the node in which 'f(i)' occurs is the intersection of the denotation of 'bottle' and the denotation of 'f(i)', after the context has supplied the values to the variables. If the context assigns to 'i' the exam the speaker has in mind when uttering the sentence, and to 'f' the extension of the relation of *being on* (relative to the world of evaluation), then the value of 'f(i)' will be *the class of entities that are on this exam*.[1] And this restricts the domain of objects we are quantifying over. According to Stanley and Szabó (2000a, 253), the semantic value of the node is given by the following meaning postulate (where 'c' above is an assignment determined by the context):

(3) $\|\text{mistake, f(i)}\|^c = \|\text{mistake}\| \cap \{x: x \in c(f)\ (c(i))\}$

The reason why the authors postulate a complex variable has to do with the phenomenon of quantified contexts (Stanley and Szabó 2000a, 250). Consider sentence (4):

(4) In most of John's classes, he fails exactly three Frenchmen.

[1] With these assignments of semantic types to the variables, Stanley and Szabó (2000a) suggest a compositional combination of their semantic values. But it is not clear what rule of composition allows for these values to be computed. None of the ones in Heim and Kratzer (1998) does. The rules of composition, as introduced in Heim and Kratzer (1998, 105), take as input the semantic value of *different* nodes, so they do not apply to elements inside a simple node. This is a potential problem for the account, but it is not one that I address in this paper.

On one reading of (4), it is true iff *for most x such that x is a class of John's, he fails exactly three Frenchmen in x.* The authors maintain that the syntactic variable approach to QDR makes the correct predictions concerning this reading of (4). The first quantifier noun phrase (QNP) makes salient a certain set of individuals that it quantifies over, those that are *classes of John's* (in the educational sense). The QNP 'three Frenchmen' is implicitly completed to *three Frenchmen in a class of John's*. Therefore, it is not sufficient to posit in the LF of the second QNP a variable of type ⟨e⟩, for individuals that are classes of John's. The individual variable 'i' cannot do the job by itself. We also need to postulate a variable that gets in the context of utterance the value *being in* (relative to the world of evaluation). This is the variable 'f', of semantic type ⟨e,⟨e,t⟩⟩.

2. The problem of the limiting case and the default value solution

The first challenge to Stanley and Szabó's account I discuss here is the following: how does the theory account for those utterances of sentences where the QNP is *complete* and so no domain restriction is needed to predict correct truth-conditions? Consider an utterance of sentence (5):

(5) Every mistake on this exam was corrected.

Suppose the utterance is such that the QNP 'every mistake on this exam' is complete. That is, the speaker does not intend to convey the thought that every *formal* mistake on this exam was corrected, or that every *spelling* mistake was corrected, or any such proposition with an extra implicit completion, but simply that every mistake on this exam (say, the salient exam) was corrected. Now, the theory relies on the context to supply a value for 'f' and 'i', but it is not clear what these values could be in the case of (5). Apparently, the context does not supply any value at all to the variables.

Bach (2000) raises this issue as an objection to the Stanley and Szabó's proposal. He considers sentence (6):

(6) All men are mortal.

He writes: "Although this is a limiting case, the value of the domain variable must still be contextually provided. Otherwise, the sentence would not express a proposition at all" (Bach 2000, 274). Stanley and Szabó (2000b) do not

address this objection in their reply to Bach's (2000) criticism. I know of no other place where they discuss this question.

A possible reply that might come to one's mind on behalf of the syntactic variable approach is that the value of 'f(i)' for the utterance of (5) is the same as for the utterance of (1) in the same context. That is, 'f(i)' stands for *the class of entities that are on this exam*, and that introduces a restriction without a difference. While this suggestion works for the case of sentence (5), it does not have a counterpart for the case of sentence (6), as here there are no corresponding plausible candidates for the values of 'f' and 'i'.

A general solution must provide default values for the variables in the case of complete quantifiers, which get us the intuitively correct truth-conditions for the utterance of the sentence. One could suggest, for instance, that the contextually determined assignment function assigns an *arbitrary* value to the variable 'i'. Indeed, there is no particular object that is salient, or in any other way relevant for the truth-conditions of (6) (in the case of (5), no other object apart from the room explicitly referred to). If the variable 'i' is to receive a value at all, even if the context does not pick out one, it can only be an arbitrary object from the domain D_e (relativized to the world of evaluation). The value of 'f' could be the extension (relative to the world of evaluation) of the property of being *either identical to or different from* an object. All individuals in the world of evaluation have the property of being *either identical to or different from* any arbitrary object. Therefore, the value of the CN would be the following:

(7) $\|\text{mistake on this exam, f(i)}\|^c = \|\text{mistake on this exam}\|^c \cap \{x = c(i) \vee x \neq c(i))\}$

This way we get the desired outcome, that of having a restriction without a difference. However, while this proposal does provide the right truth-conditions for the utterances in question, it is artificial and it very much looks like an *ad-hoc* move. It is *ad hoc*, as the only reason to postulate these values is to obtain the intuitively correct results. It is artificial in the sense that it does not seem to be the case that either the speaker who utters (5) or (6), or the hearer, entertains a thought involving the property of *being self-identical*, or a singular thought *about* an arbitrary object. This problem is especially pressing if we consider a framework of *structured propositions*. On the other hand, if we take (7) to be the contribution of the expression to the *truth-conditions* of the utterance of (5) (and make no explicit claim about the structured proposition

expressed), then this becomes an instance of the more general problem that truth-conditional semantics has with the fact that there are alternative but equivalent specifications of the truth-conditions of an utterance of a sentence. Thus, if a semantic theory assigns to an utterance of 'Snow is white' the truth-conditions: true iff *snow is white and 2 + 2 = 4*, we might suspect that something has gone wrong.

There are other options of default values that one might take the variables responsible for QDR to have. Thus, one might take the value of 'f' to be the extension of 'in' relative to the world of evaluation, and the value of 'i' to be the world of the context. However, this option will not do, because it has an undesired result, as it leads to the QNP being rigidified. An utterance of (5) will have the following truth-conditions: true iff *every mistake on this exam in* c_w *was corrected* (where c_w is the world of the context). These are intuitively incorrect truth-conditions: if we evaluate the utterance relative to a world w other than the world of the context, the truth or falsity of the utterance depends intuitively on whether the mistakes *on this exam but in the world w considered* were corrected or not.

Now, it is true that in their original article Stanley and Szabó (2000a, 252) point out that the semantic types of the variables 'f' and 'i' are set to ⟨e,⟨e,t⟩⟩ and ⟨e⟩ only as a matter of convenience, and as a "simplifying assumption". Instead, "the domains contexts provide for quantifiers are better treated as intensional entities such as *properties*, represented as functions from worlds and times to sets" (Stanley and Szabó 2000a, 252). This might help avoid the problem of rigidifying the QNP, but a problem still remains. Suppose 'f' is an *intensional* variable of type ⟨⟨⟨s,i⟩,e⟩,⟨⟨s,i⟩,⟨e,t⟩⟩⟩, and 'i' of type ⟨⟨s,i⟩,e⟩ (where 's' stands for a possible world, and 'i' stands for a time). On this account, the value of the variable 'i' is not an individual, but what is sometimes called an *individual concept*. Furthermore, it might be suggested that the default value of 'i' for the limiting case (when no QDR is needed) could be a *non-rigid* individual concept that picks out *the relevant world*. The extension of 'i' is the relevant world of evaluation. The value of 'f' is the property of *being in*. But the problem now is that a possible world is not an individual, so it cannot be the extension of 'i', as defined here. A world is a semantic value of type ⟨s⟩ (see Fintel and Heim 2011, 10). In order for the suggested solution to work, the semantic type of the variable 'i' should be ⟨⟨s,i⟩,s⟩, but that would be of no help with the cases in which we do need a substantive domain restriction.

3. The problem of the limiting case and the ambiguity solution

The above discussion indicates that it would be preferable if the predicted truth-conditions of (5) and (6) did not contain a restriction without a difference. A suggestion along these lines would be to take the LFs of (5) and (6) to carry *no* hidden variables in those cases in which the QNP is (used as) complete. In the case of sentence (5), the semantic contribution of 'mistake' to truth-conditions would be the following:

(8) $\|\text{mistake}\|^{w,c} = \lambda x_{\langle e \rangle}.x \text{ is a mistake in w}$

In those cases in which QDR is required, (e.g., the utterance of sentence (1)) the LF of the sentence does contain the hidden variables, and the semantic value for the node $[_{CN}$ mistake, f(i)] is:

(9) $\|\text{mistake, f(i)}\|^{w,c} = \lambda x_{\langle e \rangle}.x \text{ is a mistake and is c(f) (c(i)) in w}$

So, on this proposal there are two different expressions in the LF that correspond to the superficial expression 'mistake'. The interpretation function assigns to each of them its semantic value. This means that 'mistake' turns out to be *ambiguous*, instantiating a kind of lexical ambiguity, given that 'mistake' sometimes expresses the concept *mistake*, but at other times it is a context-dependent expression, expressing the concept of *mistake standing in this relation to this object*.

Now, postulating ambiguities is generally not considered to be a great way to solve problems in philosophy. Methodological considerations concerning theoretical parsimony of the kind Grice (1978, 118-119) advances immediately come to mind. "It is very much the lazy man's approach in philosophy to posit ambiguities when in trouble", reads Kripke's (1977) insightful remark. Kripke suggests a policy of caution: "Do not posit an ambiguity unless you are really forced to, unless there are really compelling theoretical or intuitive grounds to suppose that an ambiguity really is present" (Kripke 1977, 268). Are there such grounds in this case?

A theoretical consideration in favor of the ambiguity solution is that it avoids the undesirable consequence that the "restriction without a difference" solution has. But there are no intuitive grounds for favoring the ambiguity solution. On the contrary, there are intuitive considerations *against* positing ambiguity in the CN: common nouns such as 'bottle' do not seem to be ambiguous

in this way.[2,3] Now, one might find questionable the claim that intuitions about certain words being ambiguous or not are *bona fide* linguistic data for semantic theories. A semantic theory is not a study of the *intuitive* concept of meaning. Instead, it may postulate various theoretical notions of meaning or semantic value (e.g., intensions and extensions), even if these theoretical claims might be found unintuitive in some sense.

However, there are other good reasons to reject the ambiguity solution. As already mentioned, on the relevant reading of (4), the variable 'i' in 'three Frenchmen' is bound by the QNP 'most of John's classes'. But, as Breheny (2003, 63) points out, it is possible to find sentences with QNPs the domain restriction of which involves various quantificational dependencies. Consider sentence (10) (cf. Breheny 2003, 63):

(10) Some student thought no examiner would notice every mistake.

One reading of (10) is that some student x thought no examiner y would notice every mistake *made on a paper x turned in which y examines*. If we want to account for this reading in the way Stanley and Szabó do for the reading of (4) discussed above, then we need to postulate *two* complex variables of the form 'f(i)' that the QNP 'every mistake' contributes to the LF of (10). The lexical entry for 'mistake' in (9) above, which has one such complex variable, is not adequate for this case. That is, we need to introduce a new lexical entry for 'mistake', apart from the ones in (8) and (9), as follows:

---

[2] Pelletier (2003) uses similar appeals to intuitions against the Stanley and Szabó theory of QDR. The intuition is that a CN such as 'bottle' or 'student' is not context-dependent (or "contextually ambiguous", as he prefers to put it). He writes: "it seems simply unintuitive to claim that the interpretation of the same *noun* changes from context." (Pelletier 2003, 156-157)

[3] An anonymous reviewer pointed out that a defender of such an approach could respond by arguing that this could be thought of as a case of *polysemy*. Polysemy is a particular form of ambiguity, in which there are "different senses of a lexical item that bear some intuitive relationship" (Jackendorff 2002, 339). Indeed, if the proposal discussed here has any plausibility then the different senses contemplated *should* be seen as instantiating polysemy, and not homonymy, as they are systematically related. However, I take the proposal to be still problematic, as there are no intuitive grounds for this claim.

(11) $\|\text{mistake, f(i), g(j)}\|^{w,c} = \lambda x_{\langle e \rangle}.x$ is a mistake and is c(f) (c(i)) and is c(g) (c(j)) in w

This lexical entry still does not help us to account for cases which involve further dependences on quantified elements, such as in (12):

(12) Every year some student thought no examiner would notice every mistake.

On one reading of (12), it expresses the proposition that every year z some student x thought no examiner y would notice every mistake *made on a paper x turned in during z which y examines.* If Stanley and Szabó's example (4) shows that there is a complex variable in the LF of the sentence (given that it can be bound), then these readings of (11) and (12) show that there are two, and respectively three, complex variables in the LF of these sentences. With a little effort of imagination, we can build examples that involve even more dependencies of the restriction of the domain of quantification on previously introduced elements. This means that we need to postulate an indefinite number of lexical entries for CNs such as 'mistake' that differ from each other in the number of variables of the form 'f(i)' that they carry. While a language with such a lexicon is not necessarily unlearnable, as the lexical entries are introduced in a systematic way and following a pattern, this is clearly a very unattractive option.[4]

4. Other approaches to QDR that postulate hidden variables

To recap, Stanley and Szabó's syntactic variable approach to QDR gets into problems both when no QDR is required, and when the restriction needed requires that we postulate more than one complex hidden variable in the LF. In this section I argue that the two problems affect not only Stanley and Szabó's version of the hidden variable approach, but other versions as well. On von

[4] Stanley and Szabó (2000a, 232, n.16) discuss a possible ambiguity approach to QDR, but not the one considered here. On the approach they consider, a CN such as 'mistake' is multiply ambiguous, having one lexical meaning corresponding to each possible completion. They reject this option as "implausible".

Fintel's (1994, 30; 2014) proposal, the variables 'f(i)' cohabit the same node with the quantifier determiner.[5] The LF of (1) is (13), instead of (2):

(13) $[_S\ [_{DP}\ [_{DET}$ Every, f(i)] $[_{CN}$ mistake]] $[_{VP}$ was corrected]]

On the ambiguity solution to the limiting case of QDR, it is the quantifier determiner that is multiply ambiguous, having an indefinite number of lexical entries, starting with (14), (15), and so on:

(14) $\|\text{every}\|^c = \lambda g_{\langle e,t\rangle}.[\lambda h_{\langle e,t\rangle}.$ every x such that g(x) = 1 is such that h(x) = 1]

(15) $\|\text{every}\|^c = \lambda g_{\langle e,t\rangle}.[\lambda h_{\langle e,t\rangle}.$ every x such that g(x) = (c(f)(c(i)))(x) = 1 is such that h(x) = 1]

The proposal is as problematic as the similar one discussed above in relation to Stanley and Szabó's account of QDR. There are no strong intuitive of theoretical grounds for postulating a rampant ambiguity of the quantifier determiner.

Other versions of the hidden variable approach have been proposed: one of them takes 'f(i)' to occupy its own node. On this hypothesis, the LF of sentence (1) might look like this:

(16) $[_S\ [_{DP}\ [_{DET}$ Every] $[[_{CN}$ mistake] [f(i)]]] $[_{VP}$ was corrected]]

Stanley and Szabó (2000a, 255) reject this option, arguing that "one should not place such a burden on syntactic theory", but Stanley (2007, 248) explicitly embraces it.

This proposal obviously faces the same problem of the limiting case of QDR. The only significant difference with the previous cases discussed is that the ambiguity solution involves multiplying *the nodes* in which the variables occur, and that does not affect the semantic value of the CN 'mistake'. In this case the ambiguity solution does not boil down to a *lexical* ambiguity of the

---

[5] Another difference with Stanley and Szabó's approach is that von Fintel (1994) does not commit himself to any syntactic claim. He writes that the question whether the variable is present in the syntactic representation of the sentence "is an important conceptual and empirical issue that we will not be able to do justice here" (von Fintel 1994, 33).

CN, but rather to something closer to a syntactic ambiguity. The superficial form of a sentence containing a QNP has various LFs that differ from each other in the number of nodes of the form 'f(i)' to be found in the vicinity of the CN, which in turn depends on how many quantified contexts are involved in a particular reading of that sentence. But the solution is equally unattractive as the previous ones unless we are given a plausible and not ad-hoc explanation of how the variables end up in the LF. However, the proposal, as presented above, fails to do so.

A final alternative I briefly mention here is due to Pelletier (2003), on which the complex variable 'f(i)' is placed in the NP node, but not in any of its daughters. That is, the variable does not occupy its own node, but it is also does not co-habit a terminal node with another expression. For that reason, Functional Application (cf. Heim and Kratzer 1998, 105) fails to deliver the right result in this case (as it only computes the values of the terminal nodes, ignoring any other expression that is not in the terminal node). Pelletier (2003, 152) introduces a different rule of composition (call it Modified Functional Application, or MFA), as follows (where 'Det' stands for a determiner and 'N' for a noun):

$$||\text{Det N}||^c = ||\text{Det}||^c (||\text{N}||^c \cap c(f)(c(i)))$$

This is different from standard Functional Application, which we might represent here as follows:

$$||\text{Det N}||^c = ||\text{Det}||^c (||\text{N}||^c)$$

The price to pay for achieving domain restriction is the need to introduce a new rule of composition. It might not be a price too high to pay, if the account proved satisfactory. But does it? Pelletier does not discuss the two problems mentioned above, but it is easy to see how his account can deal with the problem of the limiting case: whenever the domain is implicitly restricted Det and N combine by MFA, taking into consideration the values of the variables as well; whenever the NP is complete and no domain restriction is required to derive the correct truth-conditions, Det and N combine by standard FA, thus ignoring the values of the variables (in that case it simply does not matter what default values we assign to the variables). However, the account fails to deal satisfactorily with the phenomenon of multiple quantified contexts, which requires more than one variable in the LF. So, it turns

out to be only in part better than the versions of the hidden variable approach to QDR previously discusses.

5. The variadic function approach

All the versions of the hidden variable approach discussed here face the problem of the limiting case (except Pelletier's) and the problem of multiple quantified contexts. The above discussion suggests that a "dynamic", more flexible, proposal is required, one that avoids the rampant multiplication of ambiguities, and at the same time provides the resources needed to account for the limiting case of QDR as well as for the cases of multiple quantified contexts. Fortunately, there are approaches that do allow for the needed flexibility (no variables, or more than one, up to as many as the restriction requires), as well as provide a systematic explanation of how the variables end up in the LF. In this section I briefly present the variadic function approach, and I argue that it offers a satisfactory solution to the two problems mentioned.

According to Recanati (2002, 319), a variadic function is a function from a predicate in natural language P*n* (a predicate with adicity *n*), to a predicate with a different adicity: P**n+1*, in the case of an *expansive* variadic function, and P**n-1*, in the case of a *recessive* variadic function. Thus, Recanati suggests that the prepositional phrase 'in Paris' in the sentence 'John eats in Paris' contributes a variadic function which maps the unary predicate eats (x), ascribed to John in the simple statement 'John eats', onto the binary predicate eats_in (x, l), which takes two arguments: an individual and a location. Following Recanati (2002, 321) the variadic function in the former case can be represented as follows:

$V_{\text{location: Paris}}$ (eats (John)) = eats_in (John, Paris)

The general form of an expansive variadic function could be given as follows:

$V(P(x_1,\ldots x_n)) = P^* (x_1,\ldots x_n, y)$

A variadic function creates a new predicate from a pre-existing one by changing the adicity of the latter. But it does more than that: in some cases, it also changes the content of the function, so that P and P* need not have the same content. Such is the case of the above variadic function that takes eats (x)

as input and gives eats_in (x, l) as output. Moreover, in the case of an expansive variadic function it also provides a value for the new variable that it introduces (unless the variable is bound). The role of the subscript 'location: Paris' in the above formula is precisely to indicate that the variadic function introduces a variable for location, and give to it the value Paris.[6]

Recanati introduces the apparatus of variadic functions in order to prove the invalidity of the Binding Argument, which Stanley proposes in Stanley (2000). According to Recanati, a key premise of this argument is the Binding Criterion, according to which "[a] contextual ingredient in the interpretation of a sentence S results from saturation if it can be 'bound', that is, if it can be made to vary with the values introduced by some operator prefixed to S" (Recanati 2004, 102). That is, if an ingredient dependent on the context is part of the truth-conditions of an utterance of a sentence S, and we can build a different sentence ΦS, such that the value of the contextual ingredient varies with the value of the operator Φ, then that contextual ingredient is a variable that Φ binds. If sentence S is (17), uttered with the intention to express the proposition that *John fails exactly three Frenchmen in his class*, then the contextual ingredient at issue is the nominal completion *in his class.*

(17) He fails exactly three Frenchmen.

The sentence ΦS might be (4) above, that is:

(4) In most of John's classes, he fails exactly three Frenchmen.

Now the nominal completion of the quantifier 'exactly three Frenchmen' varies with the operator 'In most of John's classes'. In accordance with the Binding Criterion, a component of the truth-conditions that varies with a certain operator is bound by that operator, and in turn, binding requires that there be a bound variable. So, the nominal completion of the quantifier 'exactly three Frenchmen' in (4) results from binding a variable that is present in the QNP. Given that the presence of the variable does not depend on whether it is bound or not, a variable of the same form must be present in the LF of (17). This variable must be saturated in order for the sentence to be interpretable, which

[6] For a more detailed presentation of variadic functions see Recanati (2002, 319-322) and Recanati (2004, 107-109).

means that the nominal completion *in his class* results from the saturation of this variable.

Recanati rejects this latter conclusion. He accepts that in (4) binding requires the presence of a variable that is bound, but not that this is evidence that the variable is present in (17) as well. The apparatus of variadic functions allows him to show how this is possible. Recanati (2004, 111) argues that not only prepositional phrases such as 'in Paris' introduce variadic functions, but also QNPs. According to Recanati (2004, 113), the quantified prepositional phrase 'In most of John's classes' contributes not only the quantifier 'for most x that are John's classes', it also contributes the variadic function 'in x'. To see how this result is obtained, let us first consider the LF of (17), before embedding it as in (4). This is (18):

(18) $[_S\ [_{NP}$ exactly three Frenchmen$]\ [\lambda_1\ [_S$ he $[_{VP}\ [_V$ fails$]\ [_{NP}\ t_1]]]]]$

(18) results from Quantifier Raising the QNP 'exactly three Frenchmen' from its position at the superficial form to the first upper S node up on the tree. This solves the mismatch problem between the verb 'fails', of type $\langle e,\langle e,t\rangle\rangle$, and the QNP 'exactly three Frenchmen', of type $\langle\langle e,t\rangle,t\rangle$. According to Heim and Kratzer (1998, 193f), the movement leaves behind a variable of type $\langle e\rangle$, called a *trace*. The trace then combines with the transitive verb, the type of which is $\langle e,\langle e,t\rangle\rangle$, thus solving the problem of the type mismatch. In order to get the right truth-conditions the trace must co-vary with the QNP. To achieve this, QR-ing the QNP also introduces a variable binder that will occupy the position in the sentence immediately after (i.e. below, in the phrase structure tree) the place where the QNP has landed, and which binds the trace. Binding requires that the binder be co-indexed with the variable it binds.

Now, the variadic function that 'In most of John's classes' introduces might be represented as follows:

$V_{\text{location: unspecified}}\ ([_S\ [_{NP}$ exactly three Frenchmen$]\ [\lambda_1\ [_S$ he $[_{VP}\ [_V$ fails$]\ [_{NP}\ t_1]]]]]$) = $[_S\ [_{NP}$ exactly three Frenchmen *in x*$]\ [\lambda_1\ [_S$ he $[_{VP}\ [_V$ fails$]\ [_{NP}\ t_1]]]]]$

The variadic function modifies the noun 'Frenchmen' to 'Frenchmen in x'. It increases its adicity: while 'Frenchmen' is of type $\langle e,t\rangle$, 'Frenchmen in' is of

type ⟨e,⟨e,t⟩⟩.[7] If we QR 'most of John's classes' to the upper S node (an optional, not mandatory move, as the QNP does not produce a type mismatch), we obtain the LF in (19):

(19) $[_S\ [_{NP}$ most of John's classes$]\ [\lambda_2\ [_S\ [_{PP}$ in $t_2]\ [_S\ [_{NP}$ exactly three Frenchmen in $x_2]\ [\lambda_1\ [_S$ he $[_{VP}\ [_V$ fails$]\ [_{NP}\ t_1]]]]]]]]$

Moreover, if we co-index the binder λ_2 introduced by QR-ing 'most of John's classes' with the variable x that the variadic function introduces, as in (19), then x gets to be bound by λ_2. The calculation of the truth-conditions of (19), which I skip here in the interest of space, gives us the desired result. As a consequence, we see that the bound variable x_2 in (4) is the contribution of the variadic function that 'In most of John's classes' introduces, and is not part of the original sentence that 'In most of John's classes' takes as argument (i.e. sentence (17), the LF of which is (18)). Variables, Recanati argues, are not part of the contribution to the LF of the sentence of the CN in the QNP the domain of which they restrict, as in Stanley and Szabó's proposal, or of some other elements in the QNP. They are the contribution of another QNP higher in the sentence, the one that binds the contextual element that restricts the domain.

Having seen how variadic functions work, I turn now to a brief discussion of the nature of the processes that generate them. According to Recanati (2002, 322), variadic functions might be either the contribution of an adjunct in the sentence (such as 'in Paris' and 'In most of John's classes') or introduced "by purely contextual means". When the variadic function is not realized phonetically, as in simple cases of QDR, its presence in the truth-conditions of the sentence is optional. In that case it is the contribution to the truth-conditions of a purely pragmatic process of "free enrichment", a non-mandatory modification of the literal content of the quantifier. However, one might depart from Recanati's view on this point, and take on board variadic functions, but not the claim that they are pragmatic mechanisms. For instance, Marti (2006) uses the

[7] This departs slightly from Recanati's (2004, 113) presentation of the output of the variadic function, where 'in x' is placed next to the S node, and takes as argument the S node. However, 'in x' has to modify the QNP 'exactly three Frenchmen', so it needs to make a direct contribution to the interpretation of this QNP. That is why I take 'in x' to modify the noun 'Frenchmen', and not the sentence node. I come back to this point in the next section.

apparatus of variadic functions, but does not take them to be the contribution to truth-conditions of a pragmatic process. Instead, Marti (2006, 141-142) suggests a purely semantic mechanism: free variables may be optionally generated in the syntax, and a variable thus generated might receive as value a contextually determined variadic function, which in turn modifies the adicity of the predicate it combines with.

As I argue below, both Recanati's and Marti's take on variadic functions have the advantage of avoiding the two problems that the hidden variable approaches to QDR discussed in the previous sections face. However, one might want to avoid any appeal to either free generation of variables or free processes such as pragmatic enrichment, which are always open to the objection of being too unconstrained and arguably ad hoc. For this reason, I prefer a semantic account, according to which quantifiers introduce a variadic function in the predicate they combine with, in virtue of their *lexical meaning*. The suggestion is that quantifiers are ambiguous, having one lexical meaning on which they introduce a variadic function, but also one on which they do not do so. Zeman (2015, 177) discusses this proposal briefly, in the context of a different, but relevantly similar debate, concerning the semantics of predicates of personal taste. He points out that worries that free pragmatic processes are too unconstrained are not sufficient reason to reject the apparatus of variadic functions *per se*, as "the variadic functions approach is in itself independent from Recanati's strong pragmatic commitments" (Zeman 2015, 178). Still, the proponent of the variadic function approach owes us an explanation of why a quantifier sometimes contributes a variadic operator and sometimes it does not. Zeman suggests that the quantifier might be responsible for introducing a variadic function in the logical form of the sentence whenever "*the truth-conditions of the uttered sentence require it*" (Zeman 2015, 178). One might further suggest, along these lines, that quantifiers have a lexical feature that allows them (without requiring that this be so on every use) to introduce a variadic function in the predicate they combine with. To avoid the charge of postulating a mechanism of ad-hoc generation of variadic functions, we could think of quantifier determiners as ambiguous: on one meaning they introduce a variadic function on the predicate they combine with, on the other they do not. For reasons that will become clear in the next section, I take the variadic function that quantifiers introduce to be able to take as argument *any* node within the predicate the quantifier combines with, and not necessarily the node of the predicate itself.

A worry might be raised at this point, as an anonymous reviewer notes. Doesn't the suggestion that quantifier determiners are ambiguous bring back all the difficulties that ambiguity approaches to QDR face, which were discussed in section 3? Indeed, some of the worries mentioned in section 3 might be reasonably raised here again. For one thing, quantifier determiners do not seem to be ambiguous. However, notice that if quantifier determiners are thought of as ambiguous, as I suggest here, this ambiguity is not multiplying beyond control. As we saw, in order to account for multiple quantified contexts, a defender of the hidden variable approach must postulate a rampant ambiguity in nouns (on Stanley and Szabó's account) or in the quantifier determiner (on von Fintel's account). The ambiguity postulated here is less problematic in this sense. So, while I acknowledge that the present proposal is not without difficulties, its comparative merits recommend it for serious consideration.

6. The two problems revisited

The virtue of the account of QDR proposed here is that it avoids the two problems discussed that the hidden variable approaches face, and it does so without postulating rampant ambiguities. Let us first look into the phenomenon of the limiting case of the QDR, i.e., when the intuitively correct truth-conditions do not require implicit completion of the nominal in the quantifier phrase. The suggestion that quantifiers are ambiguous, having one lexical meaning on which they introduce a variadic function, and one on which they do not do so, explains why in some cases the quantifier phrase is not completed in any way. The reason is that in these cases we make use of the meaning of the quantifier (for instance, 'Every mistake on this exam' in sentence (5)) that does not introduce a variadic function on the predicate it combines with ('was corrected', in the case of (5)).

Consider now the phenomenon of multiple quantified contexts. This is exemplified by sentences (10) and (12) introduced above. Consider again sentence (10):

(10) Some student thought no examiner would notice every mistake.

We can hear a reading of (10) on which it is true iff: some student x thought no examiner y would notice every mistake *made on a paper x turned in which*

y examines. The restriction of 'every mistake' involves a double quantified context. In this case we deploy the meaning of the quantifier 'some student' that introduces a variadic function on 'mistake'. We do the same for the quantifier 'no examiner'. As a result, the second variadic function returns the noun 'mistake made on a paper x turned in which y examines', the type of which is ⟨e,⟨e,⟨e,t⟩⟩⟩. To see how this is achieved step-by-step, notice that (10) results from embedding (20) under 'Some student thought'.

(20) No examiner would notice every mistake.

According to the proposal in the previous section, the QNP 'no examiner' introduces a variadic function in the predicate that it takes as argument, i.e., 'would notice every mistake'. The approach proposed in the previous section is such that the quantifier is able to introduce a variadic function on any node in the predicate that it combines with. In this case the variadic function introduced modifies the node of the CN 'mistake'. The LF of the sentence, after QR-ing the quantifiers, is the following:

(21) $[_S\ [_{NP}$ no examiner$]\ [\lambda_2\ [_S\ [_{NP}$ every mistake *on a paper which* x_2 *examines*$]\ [\lambda_1\ [_S\ t_2\ [_{VP}\ [_V$ would notice$]\ [_{NP}\ t_1]]]]]]]$

Notice that the QNP 'every mistake' has also been QR'ed from its original position to solve the type mismatch. The binder that the QNP 'no examiner' introduces (i.e., λ_2) is co-indexed with its trace (i.e., t_2) as well as with the variable x_2, and so it binds the newly introduced variable.

The next step to get (10) is to embed (20) in the phrase 'Some student thought', which again introduces a variadic function on a node that is in its scope. In particular, the variadic function modifies the noun 'mistake on a paper which x_2 examines'. The variable introduced by this variadic function (i.e., x_3) is co-indexed with the binder λ_3, which results from QR-ing 'Some student'. The result is the following LF for (10):

(22) $[_S\ [_{NP}$ Some student$]\ [\lambda_3\ [_S\ t_3\ [_{VP}\ [_V$ thought$]\ [_S\ [_{NP}$ no examiner$]\ [\lambda_2$ $[_S\ [_{NP}$ every mistake on a paper which x_2 examines *and* x_3 *turned in*$]$ $[\lambda_1\ [_S\ t_2\ [_{VP}\ [_V$ would notice$]\ [_{NP}\ t_1]]]]]]]]]]$

The computation of the truth-conditions of (22) gives us the intuitively correct result, showing that the approach to QDR based on variadic functions successfully accounts for cases of multiple quantified contexts.[8]

This analysis of (10) shows why we need the variadic function that a quantifier introduces to be able to modify a particular node within the predicate the quantifier combines with. This is part of the proposal made in the previous section (see also footnote 7). In the case of (22) we need the variadic function that 'some student' introduces to operate on the QNP 'every mistake on a paper which x_2 examines', and not directly on the whole expression the quantifier 'some student' combines with. Otherwise, the variadic function introduced would not modify an expression *within* the scope of the propositional attitude verb, and so it could not affect the nominal phrase 'every mistake'. The QNP the domain of which is restricted is embedded in an intensional context, so the restriction must be embedded as well, in order to get the correct truth-conditions.

Finally, consider simple cases of QDR such as (1):

(1) Every mistake was corrected.

In this case the quantifier is not embedded in another one responsible for introducing a variadic function that operates on 'mistake'. However, the single quantifier in (1) does introduce a variadic function. If we QR the quantifier from its position at the superficial form, we obtain (23):

(23) $[_S\ [_{NP}$ Every mistake$]\ [\lambda_1\ [_S\ [_{NP}\ t_1$ on x$]\ [_V$ was corrected$]]]]$

I suggested above that quantifiers, in virtue of their lexical meaning, might introduce variadic functions that operate on some element inside the predicate the quantifier takes as argument. In this case the variadic function operates on the trace, which can be conceived as a zero-place predicate. The value of x is given by a contextually determined assignment function, and in this case it will be *this exam*. As a result, we obtain the relevant reading of (1).

The present discussion indicates that a too rigid account of how variadic functions are introduced fails to account for the two phenomena discussed (i.e.,

[8] Alternatively, the variables x_2 and x_3 might be co-indexed with a different binder, or not co-indexed with any binder at all. These cases correspond to alternative interpretations of the sentence.

that of the limiting case of QDR and that of multiple quantified contexts). We need a flexible theory of variadic functions, both with respect to *whether* the quantifier introduces a variadic functions, and with respect to *where* it does so. For cases where no QDR is needed, we want to say that quantifiers do not introduce variadic function at all. In the case of (1)/(23) we want to say that the variadic function operates on the trace, and not on 'was corrected'. In the case of (10)/(22) the quantifier 'some student' restricts the quantifier 'every mistake', but not 'no examiner', so the variadic function it introduces must affect the interpretation of the former, but not of the latter.

The more general conclusion reached in this paper is that a theory that appeals to the apparatus of variadic functions, after several fine-grained adjustments, seems better prepared to deal with the two problems discussed (that of the limiting case of QDR and that of multiple quantified contexts) than an approach that postulates hidden variables in the LF of quantifier determiners or nouns.

References

BACH, K. (2000): Quantification, Qualification and Context: A Reply to Stanley and Szabó. *Mind and Language* 15, 262-283.

BREHENY, R. (2003): A Lexical Account of Implicit (Bound) Contextual Dependence. In: Young, R. and Zhou, Y. (eds.): *SALT XI/I*. Ithaca: Cornell University, 55-72.

COLLINS, J. (2007): Syntax, More or Less. *Mind* 116, No. 464, 805-850.

FINTEL, K. VON (1994): *Restrictions on Quantifier Domains*. Dissertation. University of Massachusetts at Amherts.

FINTEL, K. VON (2014): *Quantifier Domain Restriction*. Invited tutorial talk at Rutgers Semantics Colloquium. New Brunswick, NJ, April 4, 2014. Available at: http://mit.edu/fintel/fintel-2014-rutgers-domains.pdf.

FINTEL, K. VON and HEIM, I. (2011): *Intensional Semantics*. Retrieved from http://web.mit.edu/fintel/fintel-heim-intensional.pdf.

GRICE, P. (1978): Further Notes on Logic and Conversation. In: Cole, P. (ed.): *Syntax and Semantics: Pragmatics. Vol. 9*. Academy Press, 183-197.

HEIM, I. and KRATZER, A. (1998): *Semantics in Generative Grammar*. Oxford: Blackwell Publishers Ltd.

JACKENDOFF, R. (2002): *Foundations of Language*. Oxford: Oxford University Press.

KRATZER, A. (2004): Covert Quantifier Domain Restrictions. *Talk at the Milan Meeting*, Palazzo Feltrinelli, Gargnano.

KRIPKE, S. (1977): Speaker's Reference and Semantic Reference. *Midwest Studies in Philosophy* 2, 255-276.

MARTÍ, L. (2006): Unarticulated constituents revisited. *Linguistics and Philosophy* 29, No. 2, 135-166.

PAGIN, P. (2005): Compositionality and Context. In: Preyer, G. and Peter, G. (eds.): *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford: Oxford University Press, 303-348.

PELLETIER, F. J. (2003): Context Dependence and Compositionality. *Mind and Language* 18, No. 2, 148-161.

PUPA, F. and TROSETH, E. (2011): Syntax and Interpretation. *Mind and Language* 26, No. 2, 185-209.

RECANATI, F. (1996): Domains of Discourse. *Linguistics and Philosophy* 19, 445-475.

RECANATI, F. (2002): Unarticulated Constituents. *Linguistics and Philosophy* 25, No. 3, 299-345.

RECANATI, F. (2004): *Literal Meaning*. Cambridge: Cambridge University Press.

STANLEY, J. (2000): Context and Logical Form. *Linguistics and Philosophy* 23, 391-434.

STANLEY, J. (2002): Nominal Restriction. In: Peter, G. and Preyer, G. (eds.): *Logical Form and Language*. Oxford: Oxford University Press, 365-390.

STANLEY, J. C. and SZABÓ, Z. G. (2000a): On Quantifier Domain Restriction. *Mind and Language* 15, Nos. 2-3, 219-261.

STANLEY, J. and SZABÓ, Z. G. (2000b): Reply to Bach and Neale. *Mind and Language* 15, Nos. 2-3, 295-298.

ZEMAN, D. (2015): Relativism and Bound Predicates of Personal Taste: An Answer to Schaffer's Argument from Binding. *Dialectica* 69, No. 2, 155-183.

Organon F 23 (3) 2016: 405-416

Prior's Definition of Creative Definitions

(Sobociński-Prior-Lejewski's Discussion on the Leśniewskian Definitions)

ZUZANA RYBAŘÍKOVÁ

Department of Philosophy. Palacký University Olomouc
Křížkovského 12. 77180 Olomouc. Czech Republic
zuzka.rybarikova@gmail.com

RECEIVED: 02-02-2016 • ACCEPTED: 02-05-2016

ABSTRACT: The article introduces Prior's paper *Definitions, Rules and Axioms* which deals with Leśniewski's creative definitions. It presents the origins of Prior's paper and the discussion which is linked with its final form. Prior's aim in this paper was to present the Leśniewskian definitions in comparison with Russell's concept of definitions, demonstrating their advantages and disadvantages. The main source of Prior's knowledge about the Leśniewskian definitions were Sobociński's papers and letters, which are stored in the Bodleian library. Although the paper *Definitions, Rules and Axioms* is a unique attempt at approximating creative definitions, it contains several mistakes. Lejewski identified them in his letter to Prior and also described how they arose. Lejewski's critique was not severe, however, and Prior coped with it in the introductory page of his paper.

KEYWORDS: Arthur N. Prior – Bolesław Sobociński – Czesław Lejewski – Stanisław Leśniewski – the creative definitions.

0. Introduction

Arthur Prior's work deals with numerous problems and his papers cover a broad spectrum of logics. A number of his papers have been discussed intensively (e.g. Prior 1955a), while other have been somewhat neglected. The paper *Definitions, Rules and Axioms* belongs to the latter category, even though

it is not without interest (see Prior 1976). Prior presented an ambitious attempt to approximate the Leśniewskian definitions to logicians who were only familiar with the Russellian definitions in it, long before Rickey (1975) paper was published. This Prior's effort was quite brave but not entirely successful. It induced Lejewski's (1956) comment on Prior's paper which has the form of an autonomous paper, even though, as far as I know, Lejewski never published it.

The aim of my paper is to present an analysis of the paper *Definitions, Rules and Axioms* in order to identify the problematic points and explain the principles following Prior's correspondence with Lejewski and Sobociński. In particular, I would like to argue that at least some of them have roots in Prior's adoption of Sobociński's theory, which was not entirely in accordance with Leśniewski's theory. Furthermore, I would like to illustrate the broader problem of spreading Leśniewski's ideas among logicians, who were more familiar with Russell's system of logic, using this example. Firstly, Leśniewski's papers were not easily available (which is not the case any longer). Secondly, Leśniewski's system of logic was dissimilar to Russell's system of logic. This situation was known to Leśniewski's students and they discussed these dissimilarities intensively. They were at home, however, in Leśniewski's system and therefore sometimes failed to explain clearly all the troublesome features. This could lead logicians who based their knowledge about Leśniewski on his student's papers into misinterpretations.

Prior was acquainted with Leśniewski's system of logic through works of Łukasiewicz. Łukasiewicz used Leśniewski's axioms in his system of logic, which Prior used and praised for a certain period of his life (see, e.g., Prior 1952). It might have been Łukasiewicz who encouraged Prior to contact Sobociński and Lejewski. There is no written evidence in Łukasiewicz's letters[1] that he did so, however, thus it might also have been Prior's own idea.

In Prior's archive, we find Sobociński's and Lejewski's letters but not Prior's responses. In 1953, Prior contacted Sobociński and the following year Prior also received Lejewski's first letter. Sobociński's letters contained a detailed expression of Leśniewski's system of logic, including several proofs.

[1] Notwithstanding, Sobociński claimed in his first letter that Łukasiewicz asked him to send some offprints of his papers to Prior and Łukasiewicz (1953) mentioned Sobociński, when he claimed that Sobociński, his former student, was the editor of the Journal of Computing Systems.

Lejewski's letters are in general shorter and focused on a discussion of several features of Leśniewski's, Lejewski's or Prior's systems of logic. The Bodleian Library stores Sobociński's letters from 1953 to 1955 and then two letters from 1965, which Sobociński wrote as the editor of the Notre Dame Journal of Formal Logic. Lejewski and Prior exchanged several papers and sent comments to one other. Their letters are also more personal since they met several times and were also colleagues at Manchester University.

1. Creative definitions

Rickey (1975, 273-274) points out that Leśniewski seems to be the first philosopher to have introduced the idea of creative definitions. Leśniewski did not discuss this topic in his papers, however, but presented his ideas in his lectures and used them in his Mereology, Ontology and Protothetic. This fact is not significant for Prior. He merely read the secondary sources of Leśniewski's ideas. However, it might have affected the different understanding of the concept, which arose between Sobociński and Lejewski. Namely, Sobociński developed Protothetic during Leśniewski's lifetime but also after Leśniewski's pre-mature death (see Sobociński 1998, 70-74). In his papers as well as in his letters to Prior, he did not differentiate between Leśniewski's ideas and his inventions, even though, he clearly expressed Tarski's contribution.

Leśniewski did not consider definitions to be abbreviations. Urbaniak (2014, 152) asserts that they are more axioms than definitions. Sobociński (1953a) claimed that the new semantical category could be introduced into theory via definitions. Since the variables of the newly introduced semantical category are contained in the creative definition, the semantical category can be used in theory. Sobociński maintained that this feature made Leśniewski's system of logic a growing system, to which new semantical categories could be added. He further pointed out that these definitions had to be based on the rules of a system.

The Leśniewskian definitions are creative, but as was mentioned before, there are established rules, which have to be fulfilled. In his second letter to Prior, Sobociński (1953b) demonstrated to Prior that the contradiction, which Prior encountered, was accounted for by his violation of certain rules. The definition had to specifically have the form of equivalence, in which the position of definiens and definiendum was strictly settled. Definiens is on the

right side and definiendum is on the left.[2] The newly introduced term or entire category is consequently situated on the left side of the definition. There are also specific rules for protothetical and ontological definitions. Sobociński claimed:

> 1) In the protothetical definitions (in prototheticc, ontology, a.s.o.) the first sign of definiendum *must be* a defined constant.
> 2) In the ontological definitions (in ontology, mereology a.s.o.) the fourth sign of definiendum must be a defined constant. Any of theses constants can be followed by a row of the different pairs of parentheses. In which parentheses there are included only the variables. Each of these variables must be different from the others and all must occur in the definiens and in the main quantifier of the definiendum. (Sobociński 1953b)

The form and the use of creative definitions are also presented in Sobociński's paper *An Investigation into Prototheticc* (see Sobociński 1998) which brought additional information about the Leśniewskian definitions to Prior (cf. Prior 1955-1956, 199). Sobociński used creative definitions for introducing operators, for instance:

> If the symbolic expression 'p ⅆ q' equivalent to '~ (p ∨ q)', is introduced, so that
>
> [pq]: p ⅆ q.≡.~ (p ∨ q)
>
> becomes a valid theorem, the following theses of prototheticc can be established … (Sobociński 1998, 76-77)

Sobociński did not discuss the theory connected with creative definitions and its rules in detail in his paper. He did not even mention that he had handled creative definitions, but briefly presented that this was the way the new terms could be introduced to Prototheticc.

Although Prior based his paper about the Leśniewskian definitions on the information provided by Sobociński's letters and paper, Lejewski (1958) also wrote a paper about the Leśniewskian definitions. It was Lejewski's paper *On Implicational Definitions*, which consisted of part of his dissertation. Lejewski provided there a propositional calculus based on the implication as a sole prim-

---

[2] For a detailed expression of Leśniewski's definitions, see Miéville (2009, 29-59).

itive function by the use of implicational definitions, which have certain features of Leśniewski's creative definitions.[3] Nonetheless, this paper was published two years after Prior's *Definitions, Rules and Axioms*.

2. Definitions, Rules and Axioms

Prior discussed certain features of Leśniewski's system of logic in several of his papers (see, e.g., Prior 1952; 1953; 1955b; 1957; 1967). There is, however, a paper which deals exclusively with Leśniewski's theory of definitions, *Definitions, Rules and Axioms*. Prior introduced there two examples of the Leśniewskian definition (cf. Prior 1955-1956, 202 and 206).

He discussed the Leśniewskian definition for the first time in Prototheticc. He chose the following formula as an example:

$$\forall p \forall q \{(p \wedge q) \leftrightarrow (\forall \delta \{\delta qp \leftrightarrow [\delta p (\forall p (p \leftrightarrow p))]\})\}$$

which is Sobociński's definition of the conjunction. Discussing the form of Leśniewski's definitions, Prior argued that the essence of the theory of definition lies in the form of the definitions, i.e. that variables are bound by a universal quantifier and it is an equivalence. He claimed:

> It is, in brief, the theory that definitions are universal equivalences which we lay down in the form of axioms whenever we wish to introduce a new expression. (Prior 1955-1956, 203)

Prior additionally asserted that Leśniewski also suggested the usage of other operators in his definitions, e.g. an equivalence could be replaced by an exclusive disjunction. Prior also reformulated the definition of conjunction by the use of the exclusive disjunction. Although the second form of the definition is far more complicated than the first one, Prior admitted that it is still a permissible variant of Leśniewski's definition of conjunction.

---

[3] They are namely creative and their introduction is limited by rules. In contrast to the Leśniewskian definitions, the primitive operator which is used here is not an equivalence or an exclusive disjunction but an implication (see Lejewski 1958, 189-193).

He also discussed the ontological definition. Prior introduced the role of creative definitions in Ontology using the example of the formula (cf. Prior 1955-1956, 204-206):

I. For all *a*, *b*, *c* and *d*, if the *c* is an a and the *b* is a *c* and the *d* is a *c*, the *b* is a *d*[4]

and maintained that this example is important from a historical point of view since: "This theorem expresses the individualising force of the word '[t]he'" (Prior 1955-1956, 204). One could deduce from this:

II. For all *a*, *b* and *c*, if the *c* is an *a* and the *b* is a *c*, then the *b* is an *a*

due to the addition of a creative definition:

III. For all *a*, *b* and *c*, the *c* is a star-*ab* if and only if the *c* is an *a* and the *b* is a *c*

Prior (1955-1956, 205) emphasized that formula I cannot be proved from theorem II unless definition III is added, even though neither formula I nor theorem II contain "star-*ab*."

Prior (1955-1956, 207-208) further raised objections to this concept from the Russellian point of view. He initially pointed out that Russellians might have objected that the proper sign used in the definitions should be "=" instead of "↔" and consequently that the Russellian definitions as abbreviations has their place in the theory. They simplified the notation. The Leśniewskian definitions in contrast multiplied axioms. They did not respect the gulf between axioms and definitions and handled it with terms where the meaning was not explained.

Prior demonstrated that the creativeness of definitions can lead to a contradiction. Namely, if "the" is replaced by "every" in the formula III, then the following formula is obtained:

IV. For all *a*, *b*, *c*, every *c* is a star-*ab* if and only if, every *c* is an *a* and every *b* is a *c*.[5]

---

[4] The numerals of the formulas are different than in Prior's paper. They were changed for the sake of unity in my paper.

[5] The formula was changed in accordance with Prior's corrigenda.

This seems to be false in every possible meaning of "star-ab". Prior maintained that it is easy to prove:

V. For no *d* is it the case, that for all *a*, *b* and *c*, every *c* is a *d* if and only if every *c* is an *a* and every *b* is a *c*,

which could be reformulated as:

VI. For every *d* there is some *a*, *b* and *c* such that "Every *c* is a *d*" is not equivalent to "Every *c* is an *a* and every *b* is a *c*."

According to Prior, these three formulas cannot be all true at the same time but V and VI follow from IV and, hence, IV is contradictory.

In response to the Russellian objections, Prior (1955-1956, 208-210) claimed that if the Russellian definitions were to be consistent, they have to be formulated as rules of inference. If the definitions are neither axioms nor the rules of inference, they cannot be consistent with other postulates. This inconsistency could lead to a contradiction, as Prior demonstrated further.

Prior (1955-1956, 211-212) also coped with the objection that the meanings of the terms, which the Leśniewskian definition deals with, were not explained. He asserted that there is no settled procedure to identify the meaning of an expression. This is not just the case of Leśniewskian definitions but also Russellian definitions, axioms and theorems. There is no need, however, for a definition which could entirely cover the meaning. Prior added:

> The expressions are then 'defined' in the sense that the logician knows as much about them as he needs to know for his particular purposes; and ordinary definitions 'define' in this sense too. (Prior 1955-1956, 212)

Prior (1955-1956, 214-215) finally formulated two objections, which might arise among Russellian logicians, but which were in all probability Prior's. Firstly, he claimed that in intensional contexts, e.g. by formalizing beliefs, the Russellian type of definition is syntactically stronger than the Leśniewskian. The Russellian definitions are also more flexible in dealing with this context. He had to admit, however, that neither Leśniewski nor Russell favoured intensionality and therefore this objection was not significant for them. Secondly, Prior pointed out that the Leśniewskian definitions are not quite intuitive. Namely, Sobociński's definition of conjunction did not really correspond to the explanation of the word "and" in ordinary language. From a logical point

of view, the Leśniewskian definition is, however, more informative than Russellian. It guarantees that everything which is provable about

$$\forall\delta\ \{\delta qp \leftrightarrow [\delta p\ (\forall p\ (p \leftrightarrow p))]$$[6]

is also provable about $p \wedge q$.

To sum up, Prior attempted to introduce Leśniewskian definitions to Russellian logicians. Although he found certain objections, which could be formulated by Russell's proponents, he seemed to appreciate certain features of the Leśniewskian definitions. Notwithstanding, the two objections, the inconsistency of IV and the disadvantage of the Leśniewskian definitions in intensional logic, were not solved satisfactorily. As will be demonstrated further, Lejewski responded to both of them.

3. Lejewski's comment

Lejewski's comment consists of the commentary part and a friendly critique of Prior. Although it contained a certain criticism, Lejewski seemed to appreciate Prior's paper. The first part was meant to be a supplement to Prior's paper in which he explained features of Leśniewski's definitions, which Prior did not mention but tacitly presupposed. I entitled the second part "a friendly critique", since Lejewski himself introduced it:

> The second part includes some criticism which – I am sure – you will find not very difficult to answer. (Lejewski 1956)

Lejewski's introductory objections focused on Prior's example of the Leśniewskian definition in Protothetic. Prior chose Sobociński's definition of conjunction which Sobociński invented as the example after Leśniewski's death. It is consequently based on Sobociński's rather than Leśniewski's theory. As Lejewski stressed:

---

[6] This definition has, by no means, a correct form of the Leśniewskian definition, since, in his system of logic, every variable is to be bound (see Woleński 1989, 150). However, it is admissible as a part of previously mentioned formula $\forall p\ \forall q\ \{(p \wedge q) \leftrightarrow (\forall\delta\ \{\delta qp \leftrightarrow [\delta p\ (\forall p\ (p \leftrightarrow p))]\})\}$, where all variables are bound.

> His theory allows for definitions such as the one produced by Sobociński but it does not stipulate them. (Lejewski 1956)

Prior (1955-1956, 204) also maintained that the protothetical definitions were the Leśniewskian definitions, but that they were not creative definitions since creative definitions only appeared in Ontology but not in Protothetic. He was mistaken at that point as follows from the previous introduction of creative definitions. He had already encountered creative protothetical definitions in Sobociński's (1998) paper, but since Sobociński did not claim clearly that the system of introducing a new operator, which he used here, were creative definitions, Prior apparently did not recognise them. Lejewski was aware of Prior's mistake and wrote in his answer:

> On page 7, Professor Prior says, that 'in general we do not have 'creative' definitions in the pure theory of truth-functions (what Leśniewski called 'protothetic')'. I find it difficult to agree with this statement, because in the systems of protothetic constructed by Leśniewski one begins the deductions from the axioms by introducing definitions, which are required exclusively for their 'creative' properties. (Lejewski 1956)

It is not clear, however, why Prior maintained that there are no creative definitions in Protothetic. This claim has no support in Sobociński's papers or his letters. Sobociński in contrast introduces ontological as well as protothetical definitions. Prior might have been misled by the fact that Sobociński (e.g. 1953a), while introducing the Leśniewskian definition in Protothetic, did not write directly that they were "creative". He also did not write that they were not.

Another comment concerns a different understanding of definitions in Leśniewski's and Russell's systems of logic. Lejewski pointed out that the Leśniewskian definitions are meant to be definitions only within specific theory – there is nothing there as a definition in the absolute sense of the word. In addition, the formula which is a definition in one system could only be a theorem in another, or in a different stage of the same system. Prior's objections in the sense of intensional logic do not consequently entirely fit with the Leśniewskian definition. Leśniewski's system of logic was strictly extensional.

Lejewski additionally stressed that V and VI could not be obtained from IV. He demonstrated that V and VI are empirical statements. He specifically maintained that from Prior's formulas the formula:

VII. for all *d*, for some *a*, *b*, and *c*, it is not the case that (every *c* is a *d* if and only if (every *c* is an *a* and every *b* is a *c*))

could be obtained which is equivalent to

VIII. for some *a* and *b*, it is not the case that every *a* is a *b*

This statement is empirical and hence cannot be demonstrated in a system of logic. Statement IV, which Prior consequently criticised in his paper, was not inconsistent.

As a reaction to Lejewski's letter, Prior coped with two major objections in one page which preceded the entire paper. He admitted that protothetical definitions are also creative and suggested another statement which could demonstrate the inconsistency of Leśniewskian definitions. This statement was also, however, not in accordance with Leśniewski's Ontology, as Ontology did not assure the accessibility of empty terms. Therefore, even a reformulated proof did not actually harm the Leśniewskian definitions. Everything added to the paper is actually regularly overlooked regardless of how important a part of the paper it is. This page was not consequently included in the reprint of the paper in the book *Papers from Logics and Ethics* (see Prior 1976, 39-55).

4. Conclusion

When dealing with Leśniewskian definitions, Prior demonstrated a great deal of courage since those differed considerably from Russell's definitions, which he was more familiar with. Although it provoked Lejewski's detailed comment, Prior's mistakes were easy to correct as Lejewski had also predicted. They were mostly caused by combining Leśniewski's theory with Sobociński's later inventions and by the fact that Sobociński did not explain certain features of Leśniewski's theory to Prior which might have seemed trivial to him.

Among the objections which Prior formulated against Leśniewskian definitions, the one which neither Leśniewski nor Russell would have supplied, appeared to be crucial. From the publication of the paper *Definitions, Rules and Axioms* up to his death, Prior primarily worked with intensional logic. Leśniewski's system of logic is extensional and his definitions were adapted to

this purpose. As Prior stressed, they are disadvantageous in intensional logic. Prior did not consequently make substantial use or discuss Leśniewskian definitions further. Notwithstanding, he did not abandon Leśniewski's logic entirely.

Acknowledgments

I am grateful to professor Jan Štěpán and to two anonymous reviewers for their comments on the previous version of this paper. This work was supported by the project "Historical Solutions of Contemporary Philosophical Problems"; No. IGA_FF_2015_004 of Palacký University.

References

LEJEWSKI, C. (1956): *Letter from 23. 5. 1956 to A. N. Prior*. Unpublished manuscript stored in the Bodleian Library.

LEJEWSKI, C. (1958): On Implicational Definitions. *Studia Logica* 8, 189-205.

ŁUKASIEWICZ, J. L. (1953): *Letter from 2. 5. 1953 to A. N. Prior*. Unpublished manuscript stored in the Bodleian Library.

MIÉVILLE, D. (2009): *Introduction à l'oeuvre de S. Lesniewski: Fascicule 2: L'ontologie*. Neuchâtel: Centre de Recherches Sémiologiques Université de Neuchâtel.

PRIOR, A. N. (1952): Łukasiewicz's Symbolic logic. *Australian Journal of Philosophy* 30, 33-46.

PRIOR, A. N. (1953): On Propositions Neither Necessary Nor Impossible. *Journal of Symbolic Logic* 18, 105-108.

PRIOR, A. N. (1955a): Diodoran Modalities. *The Philosophical Quarterly* 5, 205-213.

PRIOR, A. N. (1955b): English and Ontology. *British Journal for the Philosophy of Science*, 6, 64-65.

PRIOR, A. N. (1955-1956): Definitions, Rules and Axioms. *Proceedings of the Aristotelian Society* 56, 199-216.

PRIOR, A. N. (1957): *Time and Modality.* Oxford: Clarendon Press.

PRIOR, A. N. (1967): *Past, Present and Future.* Oxford: Clarendon Press.

PRIOR, A. N. (1976): Definitions, Rules and Axioms. In: *Papers in Logic and Ethics*. Geach, P. T. and Kenny, A. J. P. (eds.), London: Duckworth, 39-55.

RICKEY, F. (1975): Creative Definitions in Propositional Calculi. *Notre Dame Journal of Formal Logic* 14, 273-293.

SOBOCIŃSKI, B. (1953a): *Letter from 16. 9. 1953 to A. N. Prior*. Unpublished manuscript stored in the Bodleian Library.

SOBOCIŃSKI, B. (1953b): *Letter from 6. 11. 1953 to A. N. Prior*. Unpublished manuscript stored in the Bodleian Library.

SOBOCIŃSKI, B. (1998): An Investigation on Protothetic. In: Srzednicki, J. T. J. and Stachniak, Z. (eds.): *Leśniewski's Systems: Protothetic*. Dordrecht: Kluwer, 69-84.

URBANIAK, R. (2014): *Leśniewski's System of Logic and Foundations of Mathematics*. Cham: Springer.

WOLEŃSKI, J. (1989): *Logic and Philosophy in the Lvov-Warsaw School*. Dordrecht: Kluwer.

Organon F 23 (3) 2016: 417-426

Joseph Rouse: *Articulating the World*
University of Chicago Press, Chicago, 2015, 423 pages

Joseph Rouse's lifelong mission appears to be to provide an adequate characterization of the role science plays in human life; and thereby to throw some new and interesting light on life itself. His conviction is that ordinary, descriptive accounts of science, revealing its methodology and its possibilities and limitations, will not do; that science is so integrally embedded within our way of life and thereby so deeply imprinted on our world that philosophy of science is inseparable from other philosophical disciplines. In particular, Rouse maintains that an account for the role of science in our lives must be deeply normative: not in the sense that it must tell us what science should do to be effective, but rather in the sense that it must reflect the fact that science is a specific outgrowth of our essentially and inherently normative practices, such that only if this deep-rooted normativity of scientific practices is taken at face value is there any real understanding of science.

Key terms around which Rouse's new book revolves are *naturalism*, *normativity*, *concepts* and *intentionality*. *Naturalism*, in Rouse's view, is a philosophical standpoint whose exponents "regard scientific understanding as relevant to all significant aspects of human life and only countenance ways of thinking and forms of life that are consistent with that understanding" (p. 3) and the core ideas of this stance Rouse finds inescapable, though he "develops these core commitments in ways that many fellow naturalists will find unfamiliar and perhaps even alien" (p. 4). *Normativity* is something that according to Rouse must be incorporated into the naturalistic framework, and this must be done so that we neither compromise naturalism, nor explain away normativity as a mere fiction. *Concepts* and *intentionality* characterize the specific ways we humans deal with our environment and with each other, distinguishing us from other creatures; conceptual understanding is the specific mode of understanding we human beings display, while intentionality characterizes our specific mode of contact with the world.

Language

One of the most illuminating motifs of Rouse's book is his sketch of the possible origins and nature of language. Many philosophers and scientists maintain that the core process behind the emergence of language is what they call "symbolic

displacement". *Perception* produces representations closely tied to the actual environment, and what is needed for *conception*, which underlies language, is the ability to unbound the representations from the environmental stimuli and let the organism put them to the kind of work that we call thinking. Although not rejecting this, Rouse wants to tell a much more complex story about the emergence of language, fearing that taking the notion of "symbolic displacement" at face value might lead us to a dangerously oversimplified picture. What plays a crucial role in Rouse's story is the concept of *niche construction* that has been introduced by several evolution theorists (see, e.g., Odling-Smee, Laland and Feldman 2003; or Kendal 2011).

Rouse stresses that though nowadays we may certainly take language to prominently serve such purposes as transmitting information and enhancing our capacities for its cognitive management (which appears to be closely connected with the "symbolic displacement"), it would be precipitate to assume that language's evolution has been driven by these very purposes from its inception. It is not self-evident that these purposes could even have been in play at the very beginning; and indeed, evolution seldom operates so transparently. The original gains driving the emergence of language might have been quite different (perhaps connected to the cohesion of social groups?); and they might have created a 'linguistic niche' for further generations to which they adapted; and when language thus became our 'second nature', it might slowly have come to gain also the purposes which we now tend to see as key.

Thus language may have started as a mere set of vocal reactions to external stimuli which as such became an integral part of the human niche so that subsequent generations of humans adapted to it by developing swift reactions to such linguistic episodes with at least part of the reactions being further linguistic episodes, the linguistic intercourse thereby growing in complexity. Thus the drive behind the evolution of language is primarily not the adaptive value of better information transfer, nor that of improved symbolic displacement, but the force of linguistic niche construction.

Here is how Rouse sees it:

> Any account of language evolution that posits direct selection for representation and information exchange must confront this difficulty head on. Such capacities would only be useful at all after the achievement of extensive representational articulation, cohesion, and precision. Its initial selective grip would be hard to understand. By contrast, the problem does not arise if articulated vocal expressiveness originally served functions other than reportorial/representational. A limited initial expressive repertoire would

> not be pointless if the initial evolutionary 'payoff' reflected needs to recognize, sustain, and coordinate larger and more amorphous social groups. (p. 119)

Hence, Rouse concludes:

> Language … initially emerges not as the product of enhanced internal capacities of a larger hominid brain but instead as a perceptually salient, developmentally effective, and selectively important behavioral dimension of the developmental and selective environment of some hominid apes. Vocal expressiveness and its behavioral integration into a transformed way of life persisted as an integral part of these organisms' ecological heritage only through its development and reproduction in each succeeding generation. (pp. 119-120)

I think Rouse's account of language emergence and evolution is both novel and persuasive.[1] It lets us escape the received wisdom that human cognition is what I would call an "inside-out" matter: that it was born in our heads (perhaps as a result of our increasing brain capacity) and language was its means of solving the problem of how to get out to be shared among individuals. His account lets us see that it may instead have been more of an "outside-in" matter: that human cognition originated in language through our increasingly complex practices and got into our heads by their internalization.

Intentionality

The concept of language, of course, is closely connected with the concept of intentionality. Hence, what does it take that we human beings display *intentionality* that our linguistic utterances and/or our mental states are *about* something? Rouse presents a useful classification of approaches to intentionality, based on two crucial distinctions. The first is a matter of distinguishing between "approaches that treat intentional or conceptual phenomena as operative-processes or as normative statuses" (p. 56). The former are those that "seek to discern features of intentional comportments that are operative in producing their directedness toward and normative accountability to their objects" (p. 56); the latter "identifies its domain with those performances and capacities that can be held normatively accountable in the right way" (p. 57). Rouse labels the approaches

---

[1] It also seems to tally with ideas that I have presented – see, especially, Peregrin (2011; 2014).

as "A" and "B" respectively, and he sides with the latter. I concur: I am convinced that it is only with the emergence of rules that a true *content* comes into the world. If you have only ways of employment of items, however complex and sophisticated, you cannot grant them more than functions; whereas once you have *rules* of how the items *should* be used, you introduce the kinds of distinction (such as that between an impossible and an improper use) necessary for making the items truly contentful (see Peregrin 2014).

Rouse's second distinction among approaches to intentionality is the distinction between those approaches which start from empty intending (which may or may not find a matching object) and those that start from intending as a relation to an object. He labels these as "1" and "2", respectively. Here Rouse sides again with the latter, and again I think he is right: allowing for intentions wholly severed from the object fulfilling them easily leads us to a solipsistic stance where you may wonder whether there is anything at all external to the intentions. Then we are prone to see the intensions as lying "inside" (a mind, a society or whatever), where the "inside" is self-standing enough not to be in any essential contact with any "outside".

Here is, then, the final classification of the approaches Rouse reaches:

> A1: operative process accounts of the constitutive structure of some domain of possible intentional comportment (e.g., the logical structure of a language, the constitutive presuppositions of a "worldview," or the essential structure of transcendental consciousness)
> A2: operative-process accounts of the causal, functional, or practical patterns of a system's interaction with its surroundings, which suffice to open a possible gap between what the system interacts with and how the system's performances "take" it to be
> B1: normative-status accounts of how the performances of a system or group of systems as a whole mostly conform to a systematically construed ideal of rationality in context, such that the goals with respect to which it would be rational are appropriately taken as authoritative for it
> B2: normative-status accounts of how a system's actual engagement with its surroundings is articulated in a way that renders it accountable to something beyond its own actual performances or those of its larger community of intentional systems (pp. 59-60)

From what was said above, it follows that I would favor the same category of approaches as Rouse, namely B2: seeing intentionality as essentially normative and essentially involved with the things that are intended. However, I have difficulty

when Rouse populates the individual compartments of this classification with approaches to be found in the literature. The distinction between the non-normative (A) approaches and the normative (B) ones fares fine: the former compartment accommodates philosophers such as Husserl, Searle, Dretske, Fodor etc., and the latter harbor those such as Brandom, McDowell, Davidson and Heidegger. The problem is with sorting out the latter into the 1 and 2 compartments. I would say that Brandom, put into the B1 cell, definitely does not belong there, and neither does Davidson. On the other hand, I am suspicious about putting McDowell into B2 rather than B1.

McDowell, as Rouse reminds us, is famously worried that our reason might end up severed from the world, "frictionlessly spinning in the void" (see McDowell 1994). Is this worry what makes him, unlike Brandom and Davidson, a good candidate for the B2 compartment? I think the converse is the case: McDowell's worry is intelligible only on the background that there is an "inside" that can be completely severed from an "outside", which would seem to me to put him into the B1 cell. Brandom, on the other hand, stresses that our linguistic practices, which give rise to primordial intentionality, cannot be thought about as severed from the things which they target. The following passage, for example, would sound like an explicit rejection of "empty intending" as the basic point:

> Discourse practices incorporate actual things. … They must not be thought of as hollow, waiting to be filled up by things; they are not thin and abstract, but as concrete as the practice of driving nails with a hammer. … According to such a construal of practices, it is wrong to contrast discursive practice with a world of facts and things outside it, modeled on the contrast between words and the things they refer to. (Brandom 1994, 332)

To preempt McDowellian worries concerning "the void" in which we can turn out to "frictionlessly spin", Brandom continues:

> Thus a demolition of semantic categories of correspondence relative to those of expression does not involve 'loss of the world' in the sense that our discursive practice is then conceived as unconstrained by how things actually are. … What is lost is only the bifurcation that makes knowledge seem to require the bridging of a gap that opens up between sayable and thinkable contents – thought of as existing self-contained on their side of the epistemic crevasse – and the worldly facts, existing on their side. (Brandom 1994, 333)

Moreover, later in the book Rouse seems to be changing his mind and relocating both McDowell and Haugeland into the B1 cell:

> McDowell, Brandom, and Haugeland each in his own way then attempts to show how conceptual understanding really does reach out to be accountable to and constrained by objects themselves. McDowell (1994) appeals to the passivity of conceptually articulated perceptual receptivity to provide the needed "friction"; Brandom (1994) claims that the game of giving and asking for reasons incorporates our causal relations with objects in perception and action; Haugeland (1998, ch. 13) argues that only an "existential commitment" to preserving an "excluded zone" of conceivable but impossible occurrences can allow objects themselves to govern what we say and do. (p. 184)

It seems, now, that all these thinkers start from an "inside" and try to "reach out" into an "outside", where genuine objects really are. I think this is unwarranted (save, perhaps, in the case of McDowell, as I have already pointed out). And as I have already ventured, I find this inadequate: I do not think that Brandom's (or, for that matter, Haugeland's or Davidson's) outlook can be construed in this way.

Conceptual normativity

Anyway, at this point the reader may become truly curious about what exactly Brandom, Haugeland and others, according to Rouse, are all lacking and what it is that he can offer. Rouse argues that "an adequate account of conceptual normativity requires the integration of biological teleology and social practice; neither alone is sufficient" (p. 161). The social practice component of conceptual normativity, I think, is straightforward: it is our taking the utterances (or perhaps, more generally, also non-linguistic antics) of others (and of our own) for correct or incorrect, ascribing them various commitments and entitlements and recognizing the potential slack between what *is* the case and what *should be* the case. However, what is the "biological teleology" component?

What may come to mind is the Millikanian version of "teleological normativity" (cf. Millikan 1984; 2004): some kind of functioning of an organism or of its organ is correct if this was the function for which the organism or organ was selected during evolution. It is, however, important to stress that this is *not* what Rouse's "biological teleology" amounts to – he explicitly distances himself from Millikan in this respect. His kind of teleology has more to do with the fact that an organism operates in an essentially goal-directed way. Rouse explains his standpoint as follows:

> We are not subjects confronting external objects but organisms living in active interchange with an environment. An organism is not a self-contained entity but a dynamic pattern of interaction with its surroundings (which include other conspecific organisms). The boundary that separates the organism proper from its surrounding environment is not the border of an entity but a component of a larger pattern of interaction that is the organism/environment complex. In the absence of appropriate interaction with a suitable environment, there is no organism because the organism dies. Death is the cessation of the constitutive ongoing pattern of interaction that is an organism making a living in its environment. After the organism's death, and especially after the extinction of its lineage, there is also no environment. An "environment" is the "belonging together" of various aspects of the organism's surroundings as collectively enabling/sustaining life. This pattern is teleological and hence normative: it has a goal, and it can succeed or fail in attaining that goal. The goal, however, is not something external to the goal-directed process but is instead the continuation of the process itself: organisms in environments are what Aristotle (1941, bk. IX) called *energeia* ("actualities"), goal-directed processes whose goal or end is present in the process itself. (pp. 186-187)

In so far as I understand the point, a biological organism is essentially goal-directed, which amounts to the "biological teleology" that constitutes the other dimension of conceptual normativity, complementing the "social practice" dimension. Thus an organism acts in a *wrong* way in so far as it does not behave in a way that fosters its inherent goals. In this way it is life itself that is an inherent source of a normativity.

However, taking life itself as yielding normativity seems to me rather problematic. Consider Pinkard's reproduction of Hegel's criticism of Kant:

> The outcome of the dialectic of "consciousness" had shown that it depended on how we were taking things, and that, in turn, raised the issue of what we might be seeking to accomplish in taking things one way as opposed to another. Thus, the issue turned on what purposes might be normatively in play (or what basic needs might have to be satisfied) in taking things one way as opposed to another. At first, it might look as if "life" itself set those purposes, and the necessary rules for judgment would be those called for by the needs of organic sustenance and reproduction. However, practical desires are themselves like sensations in cognition; they acquire a normative significance only to the extent that we confer such a significance on

> them (or, in Kant's language, only as we incorporate them into our maxims). That means that agents are never simply satisfying desires; they are satisfying a project that they have (at least implicitly) set for themselves in terms of which desires have a significance that may not correspond to their intensity. The agent, that is, has a "negative" relation to those desires, and thus the agent never simply "is" what he naturally is but "is what he is" only in terms of this potentially negative self-relation to himself – his (perhaps implicit) project for his life, not "life" itself, determining the norms by which he ranks his desires. (Pinkard 2002, 225-226)

I find this crucially illuminating: it seems to me, just like to Hegel and Pinkard, that a life can yield genuine normativity only in so far as it "has a project"; and I think that it can "have a project" only if embedded in a social network of our human type and if it participates in the network's cooperative practices. Hence, as far as I can see, it is "social practices" all the way down.

This is not to say that life itself cannot yield "its own kind" of normativity; just like evolution yields, in the Millikanian way, "its own kind" of normativity. I only insist that it is a normativity that is on a different level than that yielded by "social normativity". Once you are in the "normative space" opened up by such practices, you are at liberty to see the normativity of both evolution and life.

Science

Of course, a substantial part of Rouse's new book focuses on the nature of science and on its rule within human life. His tenet is that science is not merely one human activity among others, nor even the most important human activity; instead, it is something so deeply integrated with our way of life that it cannot be disentangled from life's other parts:

> We gain a richer and more detailed grasp of scientific understanding and scientific practice by recognizing it to be an ongoing process of niche construction. Scientific niche construction involves coordinated shifts that create new material phenomena, new patterns of talk and skillful performance, the opening of new domains of inquiry and understanding, and transformations in what is at issue and at stake in how we live our lives and understand ourselves. The sciences thereby transform the world we live in and our place and possibilities within it. In doing so, they articulate the world as conceptually intelligible. Neither merely "made up" by us nor found to have been already there, conceptual articulation is the out-

> come of new ways of interacting with our surroundings that mutually reconstitute us as organisms and the world around us as our biological environment. (p. 217)

This view of scientific practices as more of a mode of our existence than our specific activity then allows Rouse to round up his naturalistic picture of us humans within our world:

> Our discursive practices have effected a material transformation of the world and our way of life, which lets the world show itself and affect us in new ways. Our understanding of nature does not and cannot occupy an imaginary standpoint outside nature that would let us represent it as a whole in an intralinguistically articulated "image". Scientific understanding is intraworldly, partial, historically situated, and unable to transcend its own worldly involvements. Yet those involvements extend outward from scientific practices in the narrowest sense to encompass the place of scientific understanding within human life more generally. Conceptually articulated niche construction extends throughout human life. The sciences are important to us because of their integration within those broader issues, not as separate and relatively self-contained. In this respect, scientific understanding belongs within the contingencies of human history and culture. (p. 383)

In this way, Rouse presents a picture of us humans within our world that differs in many respects from that to which we are used. According to him, we are best seen not as subjects opposed to the objective world, but rather as integral parts of the objective world, which, however, must be seen as burgeoning with life; hence it is biology, rather than physics, that is crucial for our understanding ourselves. Our language, our reason and indeed our science should be seen as a natural outgrowth of the ferment of the living world, gaining its shape by means of evolution bolstered by the processes of niche construction and gene-culture co-evolution. The direction of our human life, and of our human history, is not determined to us from within: it is determined by the norms that arise out of our being the living organisms we are, and also by our being the social organisms we are. On the whole, I think that Rouse's book is duly thought-provoking – it opens new vistas on problems we thought we had already seen through.

Jaroslav Peregrin
jarda@peregrin.cz

References

BRANDOM, R. (1994): *Making It Explicit*. Cambridge (Mass.): Harvard University Press.

KENDAL, J. R. (2011): Cultural Niche Construction and Human Learning Environments: Investigating Sociocultural Perspectives. *Biological Theory* 6, 241-250.

MCDOWELL, J. (1994): *Mind and World*. Cambridge (Mass.): Harvard University Press.

MILLIKAN, R. G. (1984): *Language, Thought, and Other Biological Categories*. Cambridge (Mass.): MIT Press.

MILLIKAN, R. G. (2004): *Varieties of Meaning*. Cambridge (Mass.): MIT Press.

ODLING-SMEE, F. J., LALAND, K. L. and FELDMAN, M. W. (2003): *Niche Construction*. Princeton: Princeton University Press.

PEREGRIN, J. (2011): Creatures of Norms as Uncanny Niche Constructors. In: Hříbek, T. and Hvorecký, J. (eds.): *Knowledge, Value, Evolution*. London: College Publications, 189-198.

PEREGRIN, J. (2014): Rules as the Impetus of Cultural Evolution. *Topoi* 33, 531-545.

PEREGRIN, J. (2014): *Inferentialism: Why Rules Matter*. Basingstoke: Palgrave.

PINKARD, T. (2002): *German Philosophy 1760 – 1860: The Legacy of Idealism*. Cambridge: Cambridge University Press.