NOTES ON REFLEXIVITY

Pieter A.M. Seuren

Abstract

This paper is meant to illustrate that the deviations of language from standard logic are systematic and not due to sloppiness or lack of logical powers in human speakers. Language imposes its own constraints on the logic it operates with. The paper concentrates on the systematic fact that (binary) predicates in natural language require distinct denotations for their terms: whereas formal languages accept structures like *aRa*, where *a* denotes the same individual in both occurrences, language does not. In language, sameness of denotation leads automatically to reflexivization of the predicate in question, which, if n-ary, becomes n-1-ary. The reflexive predicate counts as different from its non-reflexive counterpart. It is shown that this constraint, the *True Binarity Principle*, or TBP, requires a *cognitive* theory of meaning, where "denotation" is the relation that holds between a term and an "address" (mental representation of an individual or set of individuals) in a discourse model. It is also shown that a number of undesirable logical consequences, including the well-known "barber paradox", are automatically eliminated from natural language by the proposed True Binarity Principle. The paper thus helps to chart the territory of logic of language, neglected for too long by logicians and linguists.

Let us begin with some simple elementary set-theory. A binary relation R in a set X can be defined as a set of pairs $\langle x,y \rangle$ such that $x \varepsilon X$, $y \varepsilon X$, and for all $z \varepsilon X$ there is some $w \varepsilon X$ such that either $\langle w,z \rangle$ or $\langle z,w \rangle \varepsilon R$. Or, in other words, R is a binary relation in X just in case all members of X figure in some pair $\langle x,y \rangle$ of R: X contains no "idle" members. Given this, it is said that R is *reflexive* just in case for every $x \varepsilon X$, $\langle x,x \rangle \varepsilon R$. R is *symmetrical* just in case, for every $x,y \varepsilon X$, if $\langle x,y \rangle \varepsilon R$, then $\langle y,x \rangle \varepsilon R$. And R is called *transitive* just in case, for every $x,y,z \varepsilon X$, if $\langle x,y \rangle \varepsilon R$ and $\langle y,z \rangle \varepsilon R$, then also $\langle x,z \rangle \varepsilon R$. If R is reflexive, symmetrical and transitive, it is called the *equivalence relation*.

It is quickly seen that if R is both symmetrical and transitive, it is also reflexive: a combination of symmetricity and transitivity is sufficient for equivalence. The proof is simple: if R is symmetrical, then for every $\langle x,y \rangle$ in R we have $\langle y,x \rangle$ in R. If R is, furthermore, transitive, then, given the symmetricity, for every $\langle x,y \rangle$ in R we have both $\langle x,x \rangle$ and $\langle y,y \rangle$ in R. Since every member of X plays a part in some pair in R, it follows that for every $x \varepsilon X$, $\langle x,x \rangle \varepsilon R$.

If X is part of some model for a formal language L we can say that the formal predicate *R* denotes the relation R, and that the name *a* denotes element $a \varepsilon X$. The formula *aRa* now means: "$\langle a,a \rangle \varepsilon R$".

The main reason for singling out reflexivity, symmetricity and transitivity as properties of binary relations lies in their set-theoretic consequences: if a relation has any of these properties, a whole train of entailments is automatically secured for expressions whose main predicate denotes that relation. These properties are thus a powerful instrument in model-theoretic semantics.

This much is clear and uncontroversial. Let us now look at natural language (in particular, English) binary predicates and see if any of these denote (binary) relations that are reflexive, symmetrical or transitive. (Actually, not wishing to be fussy, we shall simply speak of reflexive, symmetrical or transitive *predicates* whenever they denote reflexive, symmetrical or transitive *relations*, respectively.) We then notice that some English predicates are symmetrical. Thus, the predicates *meet, resemble, be like, be married to, be similar to, merge with, agree with, be identical with, equal*, to mention just a few examples, are, at least in some of their uses, symmetrical: if Harry met Susan for the first time at the birthday party I gave in 1965, then Susan met Harry for the first time at that party; if Harry is married to Susan, Susan is married to Harry, etc. As is well-known, some of these predicates allow for non-symmetrical uses. For example, I may promise you that I will meet you at the station, which is not the same as saying that you will meet me at the station. Or a daughter may resemble her mother, without it being appropriate to say that the mother resembles her daughter. Or I may equal you in ambition, without it being appropriate to say that you equal me in ambition.

(Typically, in English at any rate, symmetrical uses of predicates do not allow for passive, whereas a passive use of the predicates in question (in so far as the morphology of the predicate in question leaves room for passive) automatically implies non-symmetricity: a sentence like (1a) can be taken to be ambiguous between a symmetrical and a non-symmetrical reading, but (1b) only has the non-symmetrical reading:

(1) a. Harry met Susan at the station.
    b. Susan was met by Harry at the station.

The reason for this is probably that passivization involves a reduction of the number of arguments by one ("valency reduction"), whereas a symmetrical predicate makes no sense unless both its arguments are provided.)

Some English predicates are transitive. All comparatives, for example, are transitive: If A is larger than B, and B is larger than C, then, clearly, A is larger than C. Likewise predicates like *above* or *below*. Not, however, the predicates *to the left/right of*: although generally it will be so that if A is to the left of B, and B is to the left of C, then A is to the left of C, and if then C is to the left of D, A is to the left of D. etc., the transivity chain will continue indefinitely only if all elements are arranged in a straight line. If they are arranged in a circle, then, after a while, A will no longer be to the left of some further member, but to its right. In other words, the predicates *to the left of* and *to the right of* are not monotone. The same can be said of *above* and *below* if they are fitted into a circular geometry, but in ordinary usage they are rectilinear and monotone.

Other transitive predicates are *brother* and *sister*. Note that these are not symmetrical: John may be a brother of Susan's, but that does not make Susan a brother of John's, and analogously for *sister*. *Be a necessary condition* is transitive: if A is a necessary condition for B, and B for C, then so is A for C. Likewise predicates like *make it possible* or *enable*: if the flooding makes it possible for me to swim, and swimming makes it possible for me to escape, then the flooding makes it possible for me to escape. It will be clear that English, or any other natural language, has no shortage of transitive predicates.

Are there also predicates which are both symmetrical and transitive? Some will be inclined to answer in the affirmative. They will point at a predicate like *sibling*, which is neutral between the sexes: if A is a sibling (i.e. brother-or-sister) of B, then B is a sibling of A; and if A is a sibling of B, and B of C, then A is a sibling of C. In this vein quite a few more predicates can be found: *colleague*, *co-author*, *co-member*, *co-worker*, *coalesce with*, *be the same age as*, *live in the same village as*, and in general all predicates involving some kind of "sameness".

The problem with this affirmative answer is that whoever has a brother or a sister is then also a sibling of himself or herself: since the predicate *sibling* is symmetrical, if A is a sibling of B, B is one of A; and since this predicate is also transitive, A is then a sibling of A. Yet if I have just one sister, the number of my siblings is one, not two. Some semanticists (if they had a training in formal semantics) might reply that this is just one of those quaint features of natural language: why not simply accept that one is one's own sibling? Such a reply, however, amounts to a cover-up. First, it is simply not the case that I am my own sibling (or colleague, or co-worker, etc.) whenever I have brothers or sisters (or colleagues, or co-workers, etc.), whereas I am not when I am an only child (or employee). And secondly, even if one is prepared to accept for a fact what is not, there remains the question of *why* some facts are readily accepted, while other so-called facts systematically provoke doubts.

So, the answer ought to be in the negative: whenever a predicate is symmetrical, it is not transitive. From a purely formal point of view, this answer is correct, since the predicates in question are clearly symmetrical but their transivity breaks down on reflexive pairs. Therefore, the conclusion that they are also reflexive is not warranted, and there is no need to coerce the facts of language into any unnatural mould. Given the definitions in classical set-theory, this answer is correct. Yet it is also unsatisfactory. For the question now arises *why* it should be that these predicates are intuitively felt to be transitive whereas formally they are not. And, moreover, one wonders *why* it should be that symmetrical predicates are systematically not transitive, in natural language. Surely, there must be a general explanation for this fact.

It thus seems that a third answer is called for. This answer should do justice to the intuition of transitivity with regard to predicates of sameness, like *colleague*, *sibling*, or *contemporaneous*. Yet it should also do justice to the intuitive fact that these predicates are not reflexive. And it should provide a general explanation for the fact that symmetrical natural language predicates are not transitive in the strictly defined sense of standard set-theory.

We may thus propose to redefine the notion of transitivity for natural language:

A relation R is *L-transitive* in X just in case for every x,y,zεX, if $x \neq z$ and $\langle x,y\rangle \varepsilon R$ and $\langle y,z\rangle \varepsilon R$, then also $\langle x,z\rangle \varepsilon R$.

L-transitivity is the same as classical transitivity, except that for L-transitivity there is the extra requirement that $x \neq z$, or, equivalently, that the terms *x* and *z* denote different individuals. This effectively blocks the consequence that predicates are automatically reflexive whenever they are both symmetrical and (L-)transitive.

By recognizing L-transitivity as a significant property of binary predicates in natural language we do justice to the intuition of transitivity for predicates of sameness, while, at the same time, we save the intuition that these predicates are irreflexive. (A relation R in X is *irreflexive* just in case for every xεX, $\langle x,x\rangle \not\varepsilon R$, or: *xRx* is false for all substitutions for *x* of a name denoting an individual in X.) Yet we still lack a general principle to explain why this should be so in language. We would like to have one single explanatory principle that will account not only for the typical difference between classical transitivity and L-transitivity, but for other facts as well. (Of course, any such principle will itself call for a further explanation of why that principle should be the way it is, but we shall not, for the moment, occupy ourselves with higher planes of explanation, and limit ourselves to an attempt at formulating the best possible immediate generalization.

It seems that we are in business with the following general principle, which we shall call the *True Binarity Principle* (TBP), and which applies to all binary predicates:

True Binarity Principle:

"For all binary predicates *R*, *aRb* is uninterpretable when *a* and *b* are codenotational.

The term "uninterpretable" will be interpreted differently in different theories. In standard formal semantics it means that the function associated with R is undefined for codenotational pairs, so that no truth-value results. In discourse semantics (Fauconnier 1985; Seuren 1985) it means that an expression aRb with codenotational a and b fails to map onto a cognitive structure. The general point here is of an observational kind: any pair of codenotational terms under one predicate makes the sentence uninterpretable.

It will be clear that TBP works analogously for n-ary predicates (n>2): for any n-tuple of terms under an n-ary predicate, no two terms can be codenotational. We will, however, restrict the discussion to binary predicates.

More important is the notion of denotation. If "denotation" is taken in the sense of "reference", i.e., as a relation between a term and an element in the model (the real world if that is what we speak about), then TBP yields counterintuitive results. For if you and I are convinced that the person we both know as Ronald and the person who burgled my house last nighι are two different persons, whereas in fact Ronald is the burglar, then a sentence like *Ronald will catch the burglar* makes perfect sense, even though *Ronald* and *the burglar* are coreferential. And conversely, if you and I think that Ronald is the burglar whereas in fact he is not, then the sentence just quoted is uninterpretable. In other words, codenotation as we want it defined is independent of what is the case in the world we speak about. But it is directly dependent on the *idea* we have of that world. The crucial factor is not the world but the speaker's and hearer's *representation* of it (or of any imagined world they speak about). For this reason, TBP fits badly into any non-cognitive semantics, such as all varieties of formal model-theoretic semantics. But it is a natural part of cognitive semantic theories, such as discourse semantics.

The true Binarity Principle implies that no binary *predicate* will have in its extension a pair ⟨x,x⟩, i.e., containing the same individual twice. It does not imply that there could not be binary *relations* containing pairs of identical individuals: a binary relation may be defined by conditions that can be satisfied by one individual with respect to itself. Thus, in some model, there may be a relation designated by the predicate *love* and defined by whatever conditions hold for a truthful application of that predicate, and there may be individuals in the model that satisfy the *love*-conditions with respect to themselves. However, such pairs of identicals will, according to TBP, never be part of the set of pairs that constitutes the extension of a binary predicate. Instead, as we shall see, the individuals in question will come within the extension of a different predicate which is unary and assigns a property usually satisfied by two different individuals to one individual with respect to itself. Languages tend to have morphological and/or syntactic meańs to derive such "reflexivized" unary predicates from ordinary binary predicates (or, generally, to derive n-1-termed predicates from n-termed predicates). The standard means in English is the use of a reflexive pronoun. The contention defended here and implicit in TBP is that a binary predicate such as *love* allows for the regular derivation of a unary reflexivized predicate "self-love", expressed as *love* with a reflexive pronoun, at least in simple standard cases. But *love* and *self-love* are different predicates.

The idea that reflexivized predicates (verbs) are actually different predicates from their non-reflexivized namesakes is not novel. The (vast) literature on anaphora abounds with observations showing that reflexive pronouns are not referential but belong to the predicate, For example:

(2) a. Larry kicked himself, but Jake didn't. (McCawley 1982:137)
b. Fred broke his legs, and so did Leo.

In (2a) "Jake didn't" can only mean "Jake didn't kick himself". And in (2b), if *his* is used reflexively, i.e., as "his own", then "so did Leo" can only mean "Leo broke his (own) legs". The same phenomenon is observed in cases like:

(3) a. If you are not ashamed of yourself, I will be.
b. It's hard for Gary to understand himself, as it is for everybody.

The first of these can only be understood as saying that I will be shamed of *myself* if you are not ashamed of *yourself*. The deletion of "ashamed of myself" in the consequent clause can be accounted for only if, at some level of analysis, "be ashamed of yourself" (with the subject *you*) and "be ashamed of myself" (with the subject *I*) are identical. They are if we analyse both as a reflexivized "be self-ashamed". And (3b) can only be read as implying that *understanding oneself* is hard for everyone, not that *understanding Gary* is hard for everyone. Geach (1962:132) has the famous example:

(4) a. Only Satan pities himself.
b. Only Satan pities Satan.

and he concludes 'that here at least the reflexive pronoun is not a referring word at all'. The difference between referring pronouns and reflexives is clearly illustrated in the following pair:

(5) a. If I were you I would fall in love with myself.
b. If I were you I would fall in love with me.

Clearly, the first of these two sentences is an admonition to indulge in narcissism, whereas the second is one way, out of a multitude of ways, of making a pass at someone. In the latter sentence, the pronoun *me* is referential, and significantly not reflexive, even though the words *I* and *me* are necessarily coreferential. The point is, of course, that in this sentence they cannot be *codenotational*: after *you* and *I* have been identified in a cognitive construction expressed by the counterfactual antecedent clause, there are two cognitively distinct elements ("addresses"), the *you/I* of the cognitive construction and the *me* of the actual world, cognitively represented as such. (Although it is easy to see the point of this analysis in an intuitive way, the technical e'aboration of the mechanism involved is far from trivial. But it cannot be our purpose here to go into such technical details.)

Sometimes there is ambiguity. The following sentence:

(6) Helga thinks that she will win, and so do I.

can mean either that I think that Helga will win, in which case *she* is used referentially, or that I think that I will win, in which case *she* is used reflexively. English has no morphological marking for reflexive pronouns in dependent clauses. Some languages do, however. In the same way English fails to distinguish between reflexive and non-reflexive possessive pronouns, as was illustrated above in (2b). Some languages do have such a distinction, at least for some (usually third person) possessive pronouns.

Reflexivity is a far-reaching linguistic phenomenon, and it is not known how far it reaches. The Dutch newspaper *NRC-Handelsblad* of January 7th, 1987, contained a feature on the Swedish firm Electrolux, comparing it with the Dutch firm Philips. There the following sentence occurred (translated into English):

(7) Moreover, Philips always sends Dutch management to its foreign branches; Electrolux does so only in exceptional cases.

The context leaves no doubt that what is meant is that Electrolux sends Swedish, not Dutch, management to its foreign branches only in exceptional cases. Background knowledge is thus brought to bear to ensure the correct, reflexive, interpretation. We shall not now attempt to gauge the depths of the system underlying these phenomena. All we need here is the observation that an n-ary predicate can be reflexivized, in a very wide sense, to become an n-l-ary predicate when the therm that is eliminated is codenotational with some other term (usually the subject) or contains an attributive adjunct that is codenotational with some other term (usually the subject). This reflexivization need not be grammatically overt in all cases, yet it is semantically real.

Let us summarize, at this point. We posit that all n-ary linguistic predicates are truly n-ary. That is, each term of the n-tuple of terms constructed with an n-ary predicate in an utterance must denote a different individual. Restricting ourselves to binary predicates, we speak of the True Binarity Principle (TBP), which applies to all n-ary predicates. We have seen that the notion "denote" must not be taken in the sense of "refer", but in a cognitive sense: a term *denotes* a mental representation ("address") of an individual (whether it really exists or is just thought up). When, in an utterance, this condition is violated, the utterance is uninterpretable. This does not preclude that one can interpretably speak about cases where the conditions that hold for the applicability of a binary predicate (the predicate's *satisfaction conditions*) are, in fact, satisfied by one individual vis-à-vis itself. In such cases, however, the binary predicate is grammatically reflexivized, and counts as a different, though derived, predicate. Phenomena of reflexivization root deeply into the system of language, but we limit out immediate attention to reflexivized binary predicates.

A few things follow immediately. First, TBP automatically turns classical transitivity into L-transitivity. We may thus simply say that all classical notions with regard to relations and the predicates that stand for them are constrained by TBP. It then follows that in language a transitive predicate is simply transitive, except for reflexive pairs. A separate definition of L-transitivity is now superfluous.

Secondly, we are led to conclude that cognitive processing in so far as its output is expressed in natural language structures, is *topologically constrained*. By this we mean that, just as in the real world one individual cannot occupy two different places at the same time, cognitive representations of individuals cannot figure more than once in computational configurations. This does not necessarily apply to *variable* symbols, which do not represent individuals in some model but are mere computational tokens occurring between cognitive world-representation and their linguistic expressions. Thus, if Ronald shaves himself, we may still say that Ronald is a member of the class of individuals x such that xSx (where "S" stands for "shave"). But when we substitute denoting expressions for the variable x, only the first occurrence of x can be replaced by a denoting expression; the second occurrence will have to merge into the derived reflexivized predicate. Or, in other words, if Ronald shaves himself, he belongs to the class of self-shavers, and not to the class of Ronald-shavers, at least in so far as this fact is to be expressed linguistically. We shall come back to this point in a moment.

A few words must be said about the identity predicate *be (identical with)*. This predicate is symmetrical and transitive, and yet not reflexive, although it *symbolizes* a relation which is all three. To account for this we need some form of *cognitive, incremental* semantics (as in Fauconnier 1985, Seuren 1985). In Seuren's account (1985:427-429), the terms of the identity predicate (unless reflexivized, as in 'This is identical with itself') must denote different addresses (i.e. representations of individuals or sets of individuals) in a cognitive discourse domain. When, given some discourse domain D, a sentence is uttered with the identity predicate as its highest predicate, D is "incremented" in the sense that D will then contain also the discourse representation of the last utterance, the identity statement. The typical incremental result of identity statements is then the coalescing of the two 'old' addresses into one single new address. Or, to use Strawson's words (1974:55): "Instead of thinking of the man as operating on his knowledge-map, when his knowledge-state is changed, we may think simply of the knowledge-map as becoming changed. When he learns something from an ordinary predication, new lines inscribe themselves on his map, attached to the appropriate dot or joining two different dots. When he learns from an identity statement, the two appropriate dots approach each other and merge, or coalesce, into one, any duplicating lines merging or coalescing at the same time."

The case of the identity predicate shows again the importance of the distinction between the actual *relations* and their *predicate symbolizations*: any straightforward mapping between the two is skewed by the True Binarity Principle. It also shows the general point, repeatedly stressed above, that what counts is cognitive denotation, not factual reference.

The True Binarity Principle solves a few logical puzzles, in particular the puzzle that became known under the name of "barber paradox". We do not mean to say that TBP provides a *logical* solution to the puzzles at hand (in so far as they require one); all we maintain is that TBP provides an explanation for the *linguistic* fact that some consequences that follow in classical logical calculus are strongly counterintuitive, and that some logical paradoxes pass unnoticed in language. Our point is that, owing to TBP, the counterintuitive or paradoxical consequences may be logically valid in some, or the, classical system of logic (if no steps are taken to remove the paradoxes), but they are not valid in the logic that is operative in natural language use, this logic being constrained by certain highly specific principles, of which the True Binarity Principle is one. The logic of language thus has no need for the special provisions developed in standard logic to get rid of the paradoxes (e.g. Russell's theory of types): language has its own general principles, which have, besides their other functions, the highly practical function of avoiding paradox. In a sense, nature has developed its own safeguards against paradox.

It has been observed (Smullyan 1981:224) that in standard quantification theory the following argument is valid:

(8) a. Everyone fears Dracula.
b. Dracula fears only me.
c. Ergo: I am Dracula.

This is correct, as is easily seen (assuming that Dracula is a person and hence falls under the range of *everyone*). For (8a) entails that Dracula fears Dracula or dFd. (8b) says that dFI, and for no $x \neq I$ dFx. The only constant individual-denoting terms that will yield truth for the function dFx are those that refer to the individual referred to by d, i.e. Dracula. We now see why this argument, no matter how correct in classical calculus, fails to convince: an essential step in this argument is the entailment dFd, which follows from (8a). But under TBP this entailment simply fails to follow. For an uninterpretable sentence cannot be entailed. TBP thus appears to save intuitions. A sentence like:

(9) All the girls in her class envy Joanna.

is intuitively synonymous with "All the *other* girls in her class envy

Joanna", and does not entail that Joanna envies herself, which would be pretty odd. Under TBP this entailment is blocked, but in classical quantification theory it is not.

The consequences of this analysis (which has been presented informally and is in need of a great deal of further formal specification) are far from trivial. For one thing, it requires that variables should be interpreted as elements that stand for possible address-denoting (or, if you like, referring) *terms*, and not for the domain *elements* that they denote (or refer to). In other words, what is needed is *substitutional*, and not *objectual* quantification (no matter how painful this is for Quine's programme of elimination of particulars). For the case at hand this means that the truth-conditions of (9) say that (9) is true just in case *all sentences* of the form "*a* envies Joanna", where *a* stands for a term referring to some female class-mate of Joanna's (not, of course, to Joanna herself) are true. A statement of the truth-conditions in terms of set-inclusion (i.e., (9) is true just in case the set of all the girls in Joanna's class is included in the set of all those individuals that envy Joanna) may yield correct classical results, but for our purposes it breaks down on Joanna herself, who, though a member of her class, does not necessarily envy herself.

If an analysis along these lines is tenable, the famous "barber paradox" is eliminated. This paradox, first formulated by Russell, comes about as a result of first asserting:

(10) Jones, who is one of the villagers, shaves all the men in the village who do not shave themselves, and only those.

and then asking "Does or does not Jones shave himself?". This is a vicious question if one accepts classical quantification theory. For then, sentence (10) can be analysed as follows (x ranges over the set of villagers, S stands for *shave*, and j for *Jones*):

(11) $\forall x(\neg xSx \leftrightarrow jSx)$

This is true, in the classical style, just in case the set of those villagers who do not shave themselves is identical with the set of villagers that are shaven by Jones, or, equivalently, just in case the bi-implication $\neg xSx \leftrightarrow jSx$ is true for all assignments of x to a villager. Thus, if (10), or (11), is true, the bi-implication $\neg jSj \leftrightarrow jSj$ must be true, which means that both jSj and $\neg jSj$ should be true, which is paradoxical.

It is easily seen that under TBP the substitution of the term j for x is blocked, since the structure jSj is uninterpretable. The answer to the question "Does or does not Jones shave himself" should therefore be: "That cannot be inferred from (10): we don't know." And no paradox arises.

It will be clear that the issue at hand is a serious one and needs a much more careful formal elaboration than is given in this short note. The point of this note is to show that the deviations of natural language with respect to the formal logical analyses of mathematical logic are far from arbitrary. They are systematic and coherent, and it is therefore worth our while to investigate the underlying system. Such an investigation will inevitably lead to what logicians will feel inclined to look upon as unorthodoxy. Yet it's either unorthodoxy or a failure to understand language.

REFERENCES

Fauconnier, G. (1985), *Mental Spaces. Aspects of Meaning Construction in Natural Language*, Cambridge, Mass : MIT Press.

Geach, P.T. (1962), *Reference and Generality. An examination of Some Medieval and Modern Theories*, Ithaca, New York : Cornell University Press.

McCawley, J. D. (1982), *Thirty Million Theories of Grammar*, Chicago : The University of Chicago Press.

Seuren, P. A. M. (1985), *Discourse Semantics*, Oxford : Blackwell.

Smullyan, R. (1981), *What is the Name of this Book? The Riddle of Dracula and Other Logical Puzzles*, London : Pelican.

Strawson, P. F. (1974), *Subject and Predicate in Logic and Grammar*, London : Methuen.