ZOLTÁN GENDLER SZABÓ

COMPOSITIONALITY AS SUPERVENIENCE

What is the principle of compositionality? The short answer to this question is that it is a fundamental principle of semantics, often stated as (C):

> (C) The meaning of a complex expression is determined by the meanings of its constituents and by its structure.

As its nearly universal appearance in semantics textbooks makes clear, the principle is widely endorsed by linguists and philosophers. Not that semantic theories are usually compositional; it is often more convenient to present one's findings without worrying much about (C). What matters is the tacit commitment that in the final analysis the non-compositional wrinkles could be ironed out.[1]

Unfortunately, the harmony of opinions does not extend much beyond bare approval. It has been suggested that the principle is a trivial platitude, a significant empirical hypothesis, a useful methodological assumption, a powerful philosophical thesis in need of a deep argument, and so on. In addition, in reading the exchanges between proponents and opponents of compositionality, it is often hard to avoid the impression that those engaged in the debate have different principles in mind. So the question 'What is the principle of compositionality?' requires a somewhat longer answer.

The imprecision of (C) has been recognized by practically everyone, and attempts at tightening it up have resulted in more than a dozen formulations each of which is frequently called "the" principle of com-

[1] A nice example of the way this conviction influences semantics is the fate of Discourse Representation Theory. Kamp (1981) proposed DRT in part as a solution to problems of anaphora; he argued that the full range of data about syntactically unbound pronouns cannot be accommodated within a compositional framework. The response to this claim was overwhelming: in the subsequent years several compositional theories had been proposed which could do as much or more than the original DRT in dealing with syntactically unbound anaphora: cf. among many others, Zeevat (1989), Groenendijk and Stokhoff (1991), Muskens (1994), Dekker (1994), van Eijk and Kamp (1997).

Linguistics and Philosophy **23:** 475–505, 2000.

positionality.[2] The prevailing attitude towards these alternatives seems to be *laissez faire*: anyone is entitled to his or her compositionality and the question which of these is supposed to be the correct reading of (C) is postponed until we know more about semantics. This strategy has its merits; science often proceeds by bracketing foundational questions. Still, a bit more clarity would be useful here.

My methodology will be the *reverse* of the usual one. Instead of formulating a more precise version of (C) outright and judging the compositionality of various languages using the stipulated clarification, I will assume that, at least in simple cases, we have intuitions about whether a particular language is compositional and then I exploit these intuitions in understanding what (C) means. Semanticists are used to relying on intuitions when arguing about the meaning of various sentences; I suggest that they do the same in trying to find out how best to interpret (C).

The reading I will propose for (C) differs from its standard interpretations. My claim is not that (C) *must* be understood in the way I will suggest; it is simply that my reading fits better than others with our – not entirely pretheoretical, but still, reasonably innocent – intuitions about what it is for a language to be compositional.

In Section 1, I do some preliminary clarification and I isolate the main difficulty in understanding compositionality. In Section 2, I present three principles that are frequently used to elucidate (C) and I argue that the attempted elucidations fail. The outcome of these investigations is a better sense of what we need: (C) should be taken as a strengthening of the claim that there is a function from the meanings of constituents of a complex expression and their way of composition to the meaning of the complex expression, but the strengthening should not be achieved by demanding that the meanings of complex expressions be actually built up from the meanings of their constituents. Finally in Section 3, I suggest my own elucidation for (C) which avoids the problems that plague the standard accounts. I conclude with a few words on how my interpretation of compositionality bears on debates concerning the truth of the principle.

1. Preliminary Clarifications

There are three problems with the way (C) is worded. First, it fails to make explicit the language whose interpretation is concerned; it talks about ex-

[2] A number of these alternatives are conveniently listed in Appendix A of Janssen (1997). See also Pelletier (1994) for a discussion of the different versions of the principle of compositionality and of the way authors tend to oscillate among formulations of varying strength.

pressions and meanings in general. Second, it employs the terms 'meaning' and 'structure', which are open to a bewildering array of interpretations. (The same holds for 'constituent', but one might hope that fixing what we mean by 'structure' would take care of this additional difficulty.) And finally, (C) talks about determination, leaving what it is for something to determine something else completely unspecified. In this section, I will discuss the first two of these problems in two separate subsections. The third one, which I think is the real difficulty with (C), will be discussed in the rest of the paper.

1.1. *Possible Human Languages*

Resolving the first problem requires us to supply the language variables missing from (C), giving us (C′).

(C′) For every complex expression e in L, the meaning of e in L is determined by the meanings of the constituents of e in L and by the structure of e in L.

We can think of the variable L as being free, or as being bound by a tacit universal quantifier. If L is free in (C′), then its value is given by the context of utterance, and so the principle governs the semantics of a particular language under discussion. This is clearly the wrong way to think of compositionality. For the *evidence* we have in support of the principle is of a very general sort and the generality of the thesis should match the generality of the evidence. However one would try to justify the claim that the meaning of the phrase 'gray elephant' in English is determined compositionally, it better be an argument general enough to apply to phrases in Turkish and Swahili as well. The most common considerations in favor of compositionality – the learnability of languages, their systematicity, the ability of speakers to understand and produce complex expressions they never heard before, etc. – have nothing to do with the details of the syntax and the semantics of particular constructions in particular languages.[3] And if this is so, we should not construe the principle of compositionality in a language-specific manner. To do so would be to mismatch *what* we say and *why* we say it; it would be akin to construing the principle of gravity as being exclusively about, say, cubical objects made of wood.

[3] Cf. Grandy (1990), p. 557 " . . . in spite of the fact that we have no adequate semantics for any natural language we feel that there must be compositional semantics for all natural languages". For skeptical arguments against the sufficiency of this type of evidence for establishing compositionality, see Hintikka (1981) and Schiffer (1987); for replies see Partee (1984), Partee (1988), and Janssen (1997).

Having excluded the free variable interpretation for (C′), we must opt for the other one, according to which (C) tacitly quantifies over languages. We should think of the claim that a certain semantic theory of English is compositional as a way of saying that the theory presents English as a compositional language, and we should think of the claim that English is a compositional language as a way of saying English is one of the languages within the domain of the tacit quantifier in the principle of compositionality.

This raises the question what languages (C) is quantifying over. Construed as an unrestricted quantification, the principle would be an obvious falsehood. Not *all* languages are compositional: surely a hypothetical language where the meanings of complex expressions are influenced by the weather, while their structure and the meanings of their constituents are not, would be non-compositional *by definition.*[4]

(C) is a thesis about human languages, like Afghan, Estonian, or Spanish, but not about artificial languages, like the language of set theory, secret codes, or DOS. Languages that evolved naturally, but are more or less dead tongues by now, like Kamas, Latin or Syriac, are intended to be included, as are languages that will evolve from those languages now spoken. Languages made for limited practical purposes are not at issue, even if they happen to conform to the principle, for we could have crafted other, non-compositional languages in their stead which would have done the same job, although perhaps less conveniently.

I think the best way to characterize the domain of the tacit universal quantification over languages in the principle of compositionality is to say that the thesis concerns *all possible human languages*[5]. I suggest that we understand (C) as a somewhat contracted way of expressing that which is made explicit by (C″):

[4] Of course, one might then *redefine* what the meanings of lexical items are and show that, given the new meanings, the meanings of complex expressions depend exclusively on the meanings of their parts and the way those parts are combined. But it seems to me that this would be an altogether *different* language. By language, we usually mean *interpreted* language, so if the lexical items have different meaning, the language cannot be the same. If one uses a purely syntactic criterion for identifying languages and if one places no constraints whatsoever on what sort of things meanings are, one might trivialize (C) altogether. I discuss this issue further in footnote 6.

[5] Possible human languages are also called ‘natural languages’. I use my less conventional label to emphasize that some of these have never been and will never be spoken by anyone anywhere.

(C″) For every possible human language L and for every complex expression e in L, the meaning of e in L is determined by the meanings of the constituents of e in L and by the structure of e in L.

In what follows, I will always understand (C) as (C″), even if for the sake of convenience I suppress the quantification over possible human languages.

One might wonder what makes a language a possible human language. All I can propose here is a necessary condition: A possible human language must be at least (i) a language suitable for the expression and communication of a wide range of thoughts, and (ii) a language that can be learned by human beings under normal social conditions as a first language. The language of pure set theory and the language of traffic signs are not possible human languages because they violate (i). One cannot use them to say that one has a headache. A language whose expressions are each at least a hundred phonemes long, or a language with only two phonemes are not possible human languages, because they violate (ii). The expressions of the first language would be too hard to keep in mind, while the expressions of the other would be too easy to confuse with one another.

Exactly which languages are possible human languages is, of course, an open question. If one is given a detailed description of a linguistic phenomenon, it may be very hard to tell whether such a phenomenon could occur in a human language. (Could there be a human language without adjectives or adverbs? Could there be a human language without ambiguous words?) But the difficulty does not differ in kind from the difficulty we have in trying to decide whether a particular event of which we have a detailed description is physically possible. (Could there be a universe without one of the four fundamental forces? Could there be non carbon-based life?) The notion of a possible human language is not murkier than the notion of a physically possible world.

1.2. *Meaning and Structure*

Much of the elusiveness of the principle of compositionality derives from the two crucial concepts it involves: *meaning* and *structure*. These uncertainties are, alas, irremediable. This is so partly because of the complexity of these notions, but more importantly, because of what the principle of compositionality is supposed to *do* in our semantic theorizing. (C) is supposed to be a tie-breaker, something we can use in debates between alternative semantic theories which cover roughly the same empirical ground. If one of the theories is compositional, while the other is not, we are told to choose the former over the latter. But in order for (C) to be

applied in this regulatory manner, it cannot have one specific conception of meaning or structure built into it.

Of course, not everything goes. For example, to assign Gödel numbers as meanings to expressions of a language, or to characterize the structure of its sentences simply by the linear order of the words they contain will clearly not do. Nevertheless, to maintain as much generality as possible, we should set only minimal constraints. I propose the following two: [6]

(I) The meaning of an expression is something that plays a significant role in our understanding of the expression.

(II) The structure of a complex expression is the syntactic manner in which the expression is combined from simpler expressions.

(I) is an extremely liberal principle which allows for a great variation in the sorts of entities one might want to associate with the expressions of a language as their 'meaning'. For example, in the case of the noun 'house', any of the following will do: the set of all actual houses, the set of houses in all possible worlds, the property of being a house, a function from contextual indices to the set of those things that are houses relative to those indices, etc. Whether in the end we should say that any of these is *the* meaning of 'house,' or we should work with multiple notions of meaning depends on a variety of philosophical and perhaps psychological considerations.

(II) relegates to syntax questions concerning what the ultimate constituents of expressions are, and what mechanisms we use to put them together. Since what syntacticians mean by structure is often not very structure-like – for example, it may be a derivational history – this amounts to a liberal use of the term 'structure'. And since there is a lot of variation in detail among syntactic theories, in applying (II), one should proceed with

[6] If one places no constraints on what counts as meaning and what counts as structure, (C) becomes almost completely vacuous. Zadrozny (1994) proves that given a set S of strings generated from an arbitrary alphabet via concatenation and a meaning function m which assigns the members of an arbitrary set M to the members of S, we can construct a new meaning function μ such that for all $s, t \in S$ $\mu(s.t) = \mu(s)\mu(t)$) and $\mu(s)(s) = m(s)$. (The values of μ are functions whose values are defined using the so-called Solution Lemma, which is provable in the theory of non-wellfounded sets.) Let us accept for a moment that fulfillment of the first of these conditions guarantees compositionality. (I will actually argue against this claim in Section 2.3.) Then the theorem shows that an arbitrary meaning-assignment can be imitated compositionally. Of course, if we reject that complex expressions of human languages could have such a primitive syntax and such fanciful meanings, we can deny that μ is a meaning assignment. So, if we accept (I) and (II), Zadrozny's result poses no direct threat to the empirical significance of compositionality. For a similar assessment, see Kazmi and Pelletier (1998) and Westeståhl (1998).

caution. For example, $[_{NP}[_{Det}\text{every}][_{N}\text{man}]]$, $[_{NP}[_{Det}\text{every}][_{N'}[_{N}\text{man}]]]$, or $[_{DP}[_{D'}[_{D}\text{every}][_{NP}[_{N'}[_{N}\text{man}]]]]]$, should all count as permissible representations of the syntactic structure of 'every man'. Whether in the end we should say that any of these correctly captures *the* syntactic structure of phrase depends, again, on a variety of syntactic and perhaps psychological considerations.[7]

In what follows, I will understand (C) in this schematic manner. It says that meaning and structure – whatever exactly they might be within the bounds of the constraints (I) and (II) – are interconnected in a specific manner: namely, in possible human languages the meanings of complex expressions are determined by the meanings of their constituents and by their structure.

Having decided to interpret the principle of compositionality as a claim about all possible human languages and having acknowledged that its full content depends on further theoretical decisions about meaning and structure, it may well seem that the task of clarification is done. According to (C), in all possible human languages the meaning of complex expressions – phrases, clauses and sentences – depends on two kinds of facts. Syntax tells us what the lexical constituents – words and smaller morphemes – occurring in a complex expression are and how they are combined; and the lexicon tells us what those constituents mean. According to the principle, such syntactic and semantic features of a complex expression are jointly sufficient to determine the meaning of that expression.

This sounds crystal clear, but it is not. For despite the naturalness of wording, it is not obvious what it means to say that certain syntactic and semantic features of an expression jointly *determine* what that expression means.

[7] Higginbotham (1986) distinguishes among three sorts of questions that arise in the course of investigating the semantic nature of a linguistic construction. Questions of the first type concern the nature of semantic values: what sort of objects should one assign to these constructions, or to their parts? Questions of the second type concern the syntax of the construction: what categories do the basic elements of the construction belong to and how are they combined? Finally, questions of the third type concern how, given certain assumptions of how questions of the first and second type should be answered, semantic values get assigned to the constructions under investigation. For Higginbotham, only questions of the third type are questions for semantics proper. In trying to minimize the commitments about what 'meaning' and 'structure' mean in (C), I have made an attempt to keep compositionality within the domain of semantics proper, in Higginbotham's sense.

2. ALTERNATIVE FORMULATIONS

Repeating the same words too often is bad style; alternative ways of expressing the same thought can often elucidate one's original meaning. But the variations may also blur important distinctions. The following formulations of the principle of compositionality all come from a recent book on logic and language:

(1) The principle of compositionality [...] requires that the meaning (and thus the truth value) of a composite sentence depends only on the meanings (truth values) of the sentences of which it is composed.[8]

(2) [...] the interpretation of a complex expression is a *function* of the interpretations of its parts. This is the principle of compositionality of meaning, also referred to as 'Frege's principle.'[9]

(3) [...] the meaning of a composite expression must be built up from the meanings of its composite parts. This principle, which is generally attributed to Frege, is known as *the principle of compositionality of meaning.*[10]

(4) If two expressions have the same reference, then substitution of one for the other in a third expression does not change its reference. If two expressions have the same sense, then substitution of one for the other in a third expression does not change its sense. [...] The two principles are also known as *principles of compositionality*, of reference and sense, respectively.[11]

I think none of these principles captures exactly the intuitive meaning of (C). Some of the differences are trivial, others are rather instructive. To see things more clearly, first a few cosmetic changes are necessary: (1) talks only about sentences, not complex expressions in general; (4) comprises two principles, which concern different notions of meaning; (2) and (3) talk about 'parts', which are presumably the constituents of the complex

[8] Gamut (1991), Vol. 1, p. 28.
[9] Gamut (1991), Vol. 2, p. 140.
[10] Gamut (1991), Vol. 1, pp. 5–6.
[11] Gamut (1991), Vol. 2, pp. 11–2.

expression; (2) mentions the 'interpretation' of expressions, which must be the assignment of meanings to those expressions; and finally, (1) and (2) forget to mention the structure of complex expressions as an additional feature in fixing their meaning. With requisite changes in place and imposing certain stylistic uniformity, we get the following four principles:

(1′) The meaning of a complex expression depends only on the meanings of its constituents and on its structure.

(2′) The meaning of a complex expression is a function of the meanings of its constituents and of its structure.

(3′) The meaning of a complex expression is built up from the meanings of its constituents.

(4′) If two expressions have the same meaning, then substitution of one for the other in a third expression does not change the meaning of the third expression.

(1′) is a stylistic variant of (C): if the meaning of a complex expression *depends only* on the meanings of its constituents and on its structure, then these factors *determine* the meaning of the complex expression. The other three exemplify what I take to be the most important attempts to present the content of (C) in a more explicit form. (4′) is a principle that tries to elucidate (C) *indirectly* without talking about the nature of the determination relation that holds between the meaning of complex expressions on the one hand, and the meanings of their constituents and their structure on the other. (2′) and (3′) are *direct* elucidations. (2′) tells us that this determination is functional, (3′) says that the determination is akin to a part-whole relation: meanings are constructed from simpler meanings.

Before I turn to further discussion of these principles, I want to mention quickly one crucial ambiguity in (C) that I will *not* talk about. According to the most natural reading of the principle of compositionality, the constituents on whose meaning the meaning of the whole depends can themselves be complex. There is a weaker reading of the principle, according to which the constituents in question must be the ultimate ones, and hence they are all simple. This latter reading allows for the possibility that in interpreting a complex expression, we have to take into account not only the meanings of its immediate constituents and the last step of its syntactic construction, but also how the meanings of the immediate constituents were determined by their simpler parts. In my discussion – as is customary in formal semantics – I will assume the stronger reading.[12]

[12] Here is an example how the difference between these two readings might matter. According to some conceptions of meaning, all tautologies mean the same thing. Those who

2.1. *The Function Principle*

According to (2′), the meaning of a complex expression is functionally determined by the meanings of its constituents and by its structure. I will call this *the function principle* (F).

(F) The meaning of a complex expression is a function of the meanings of its constituents and of its structure.[13]

The claim that the As are a function of the Bs means that there is a one-many relation f between the As and the Bs, i.e., that there are no Bs that are f-related to more than one A. So, one can rephrase (F) by saying that expressions combined in the same way from synonymous constituents are themselves synonymous.[14]

It is obvious that this is a consequence of (C): if the meaning of a complex expression depends only on the meanings of its constituents and on its structure, then if we have two complex expressions sharing the same structure and having synonymous constituents, they must mean the same. The problem is with the converse claim; I will argue that (F) does not entail (C).

Assume for the sake of argument that English is compositional. (If this is too much to ask, assume that in the argument that follows, 'English' refers to a large compositional fragment of English.) Let Crypto-English be a language very similar to English, containing the same expressions with almost the same interpretation. The meanings of the expressions in

work with such a conception will have to deny that languages containing propositional attitude verbs are compositional in the stronger sense, since sentences attributing to someone a belief that some tautology is true are clearly not all synonymous. Still, compositionality in the weaker sense can be maintained because, arguably, if T and T' are tautologies such that one but not the other is believed to be true by someone, T and T' will have different structure.

[13] Compositionality is often formulated as the so-called *rule-to-rule* principle: Complex expressions can be assigned meaning by assigning meaning to the basic expressions and by giving a semantic parallel to the syntactic construction steps. The idea behind the rule-to-rule principle is that for each syntactic rule that specifies a way of combining certain sorts of expressions into a complex expression, there is a corresponding semantic rule which specifies a function that assigns to the meanings of the constituents the meaning of the complex expression we get when we apply the syntactic rule to them. (For more formal versions of the principle, see Montague (1970), pp. 231–3, Janssen (1983), p. 25, and Janssen (1997), pp. 447–53.) Assuming that the 'constituents' mentioned in (F) can be themselves complex (see my remarks in the last paragraph before 2.1.), the rule-to-rule principle is equivalent to (F).

[14] Of course, one can interpret the expression 'function of' in (F) differently. Throughout the following discussion, I will assume this strict mathematical understanding of the claim that something is a function of something else.

Crypto-English are obtained from the corresponding meanings in English via a permutation p. This permutation leaves every meaning as it is in English, except that it interchanges the meaning of sentences synonymous with (5) and the meaning of sentences synonymous with (6).

(5) Elephants are gray.

(6) Julius Caesar was murdered on the ides of March.

I argue that Crypto-English does not violate (F), but violates (C). Hence, the example refutes the claim that (F) implies (C).

Let me start with the claim that Crypto-English does not violate (F). Suppose it does; then there are two complex expressions c_1 and c_2 that are not synonyms in Crypto-English despite the fact that they were constructed in the same way from constituents that are synonyms in Crypto-English. Since it is obvious from the definition of p that two expressions are synonyms in English just in case they are synonyms in Crypto-English, we can conclude that c_1 and c_2 are not synonyms in English despite the fact that they were constructed in the same way from constituents that are synonyms in English. This implies that English violates (F), which contradicts our assumption that English is compositional. So the initial assumption is false and Crypto-English does not violate (F). Q.E.D.

Now let us consider the claim that Crypto-English violates (C). This can also be proven by a *reductio*. Suppose Crypto-English is compositional. Then the meaning of (5) in Crypto-English is determined by its structure and by the meanings its constituents in Crypto-English. We assumed that English is compositional, so the meaning of (5) in English is determined by its structure and the meanings of its constituents in English. But by the definition of p, the meanings of subsentential constituents are the same in English and Crypto-English. Consequently, the meaning of (5) is determined by the same factors in both languages, and hence, (5) has the same meaning in both languages. Since by our assumptions what (5) means in Crypto-English just what (6) does in English, this entails that (5) and (6) are synonyms in English, which is absurd.[15] So the initial assumption is false and Crypto-English violates (C). Q.E.D.

One might object that this argument is question-begging. All I have shown is that the meaning of (5) is determined through *different* functions in Eng-

[15] What if one uses an extremely coarse conception of meaning, according to which (5) and (6) *are* synonyms? (For example, for certain purposes, one might identify the meanings of sentences with their truth-values.) In that case, one should replace these sentences in the definition of Crypto-English with different ones that have different meanings according to this extremely coarse conception of meaning.

lish and in Crypto-English. How would that entail a violation of (C)? What I wanted to prove was that (F) does not entail (C), but here I seem to rely on the assumption that (C) demands something beyond there being *some* function in Crypto-English which assigns meanings to complex expressions in terms of the meanings of their constituents and their structure.

But I don't think that the argument is circular. What I rely on is not the conclusion, it is the ordinary meaning of the word 'determine'. Consider the following example. I take it that if we found two people with the same genes and different eye colors, that would amount to a refutation of the thesis that genes determine eye color. Similarly, if there are two expressions with different meanings but the same structure and constituents with the same meaning, that amounts to a refutation of the claim that the meaning of both of these expressions is determined compositionally.

One might still object that the analogy is unfair. If there are two individuals, both in the species *homo sapiens*, who have identical genes but different eye colors then it is false that genes determine eye color *in humans*. So if we found two expressions *in the same language* with synonymous constituents and identical structure but different meanings, compositionality *for that language* would be refuted. But in the example, the crucial expressions belong to different languages. So, nothing defeats the hypothesis that both English and Crypto-English are compositional; it is just that the meanings of complex expressions depend on the meanings of their constituents and on their structure in a *different way* in these languages.

But couldn't the thesis that genes determine eye color be refuted differently? Suppose we have a chimpanzee and a human being such that all the genes which can conceivably play any role in fixing eye color are the same in these two creatures. And suppose that the chimpanzee has black eyes, while the human being has blue eyes. What should we conclude? It seems to me that the correct conclusion is that at least in one of these species genes do not determine eye color. To say that eye color is determined in both species by the exact same genes in different ways is obfuscation. For there must be something about these creatures that explains the difference in eye color, and if the genes themselves can't do this then we have to look for other factors. Analogously, there must be something about the English sentence 'Elephants are gray' and the Crypto-English sentence 'Elephants are gray' which can account for the fact that they are not synonymous. If the meanings of their constituents and their structure won't do the job, we must conclude that there is some hidden factor which plays a role in fixing the meaning of at least one of these sentences. So, if the meaning of (5)

is determined compositionally in English, it cannot be so determined in Crypto-English.

As I mentioned in the introductory part of the paper, my aim is not to convince my reader that (C) *must* be interpreted in a particular way. I suppose one could use the word 'determine' is such a way that finding a chimpanzee and a human being with different eye colors but identical eye-color determining genes would not threaten the thesis that genes determine eye color in both species. But I strongly suspect that this is not the ordinary understanding of the word 'determine'.

But there is another objection against my argument. One might claim that (5) is not a *genuine* complex expression, but rather an *idiom* of Crypto-English. Idioms are expressions whose syntactic complexity is semantically irrelevant. And if (5) is a simple expression in Crypto-English, the fact that its meaning is not determined compositionally does not entail that Crypto-English is not compositional.[16]

The short answer to this objection is that (5) cannot be an idiom of Crypto-English, for the meanings of its constituents are in fact *relevant* for determining its meaning in that language. One can take the meanings of the constituents of (5) in Crypto-English, apply the inverse of p, use the syntax of English, apply p, and obtain the meaning of 'Elephants are gray' in Crypto-English, i.e., that Julius Caesar was murdered on the ides of March. Crypto-English can be interpreted via English using p and its inverse as *translation functions*.

One might object that my characterization of idioms is unfair: it is not a necessary feature of idioms that their constituents play absolutely no role in determining their meaning. But even this is granted, understanding an idiom requires *specific* knowledge about the expression beyond its structure and the meaning of its constituents. This explains why we have to learn idioms one by one, which in turn explains why there are relatively few idioms in languages we can learn. However, one could define another language, Crypto-English$^{\infty}$ using a permutation function p^{∞} which changes the meanings of infinitely many English sentences.

Let us say, for example, that the permutation function p^{∞} leaves the meaning of every expression unaltered, except that it switches the meaning of sentences synonymous with 'There are two times *N* apples on the table' with the meaning of sentences synonymous with 'There are two times *N*

[16] One occasionally encounters the claim that the presence of idioms in a language shows that the language is not compositional. That is not necessarily the case. As long as it is theoretically acceptable to treat idioms as simple expressions with semantically spurious syntactic complexity, idioms pose no danger to compositionality. For a detailed discussion of how to accommodate idioms within a compositional theory see Nunberg, Sag and Wasow (1994).

plus one apples on the table', where *N* is a schematic letter for the numeral whose denotation is the natural number *n*. Speakers of English can easily understand Crypto-English$^{\infty}$. (In fact, you have already mastered it by the time you are reading this sentence.) Since we do not think that one can pick up (and so quickly!) a language with infinitely many idioms, we are forced to reject the idea that all sentences whose meanings are not mapped to themselves by p^{∞} are idioms of Crypto-English$^{\infty}$. By parity of reasoning, we should reject the suggestion that (5) is an idiom of Crypto-English: if such a suggestion does not work in general against the type of argument I made, why try it in the specific case of Crypto-English?

The conclusion of this subsection is that (F) is weaker than (C). By saying that the meaning of complex expressions is determined by the meanings of its constituents and by its structure we say *more* than that there is a function from the latter to the former. The question is, how much more?

2.2. *The Building Principle*

Frege held the view that "corresponding to the whole-part relation of a thought and its parts we have, by and large, the same relation for the sentence and its parts."[17] Thoughts have a structure similar to the sentences which express them: as a sentence is built up from basic constituents, the sense of the sentence is constructed from the senses of the constituents. Russellian propositions are similar in this respect to Fregean thoughts: they are complex entities whose structure corresponds more or less to the structure of sentences that express them. The difference is only that Russellian propositions are built from the referents of words, not from their senses.

Some linguists and philosophers hold Fregean or Russellian views about the nature of meanings of complex expressions and endorse (3′).[18] I will call (3′) the *building principle*:

(B) The meaning of a complex expression is built up from the meanings of its constituents.

This is a thesis that implies the principle of compositionality: if the meaning of a complex expression is built up from the meanings of its

[17] Frege (1919), p. 255. Cf. also Frege (1892), p. 193, Frege (1914), p. 225, Frege (1906b), p. 192 and Frege (1923), p. 390.

[18] For example Cresswell (1985), p. 25–31 and Jackendoff (1983), p. 76. The commitments are often more complicated. For example, Kaplan (1977) distinguishes between two notions of meaning: the content and the character. The content of a sentence is a complex built from the referents of the constituent expressions, while the character is a function that assigns to each context of utterance the content of the sentence uttered in that context. (B) holds for the first, but not the second of these notions of meaning.

constituents, then surely what the meanings of the constituents are and how they are put together determines what the meaning of the complex is. However, the building principle requires something beyond compositionality. To say that the meaning of a complex expression is built up from other meanings implies – assuming the ordinary understanding of what it is to build up something from other things – that the meaning of a complex expression is a complex entity.[19]

To see that (C) does not entail (B), consider an intensional semantics where the meanings of sentences are sets of possible worlds. Such a semantics can be compositional, but it violates the building principle, since there is nothing in a set of possible worlds that could directly correspond to meanings of constituents. Sets of possible worlds are unstructured entities, and as such, they cannot be meanings according to (B).[20]

One implication of the building principle is that the constitution of meanings fixes their identity, i.e., that sameness of meaning entails sameness of structure. So, if English conforms to (B), we have to conclude either that 'The color of elephants is gray' and 'Elephants have gray color' differ in meaning, or that they have the same structure. Such conclusion is not forced upon us by compositionality itself, which is good, for the claim is not particularly plausible.

Consider the analogy of chemical compounds, to which Frege often appealed. He noted that just as certain atoms can enter into chemical reactions where they are combined in complex configurations, so too can senses of words be put together to build senses of phrases and sentences.[21] But the chemical analogy can be turned against Frege. Radicals are complexes built up from unique collections of simpler constituents, but their corresponding chemical properties are not. So while it is true that nothing could be hydroxyl (HO^{-1}) without being composed of a hydrogen nucleus and an oxygen atom, it is also true that there are many different radicals – built up

[19] When I interpret (B) in this manner, I take the building metaphor very seriously. It is quite clear that many authors who have formulated the principle of compositionality as (B) do not do so. Perhaps all they intend to express by (B) is the claim that there is a function from the meanings of parts to the meanings of wholes. It is sometimes suggested (e.g., Salmon (1989), p. 438) that Frege himself intended only this much. In 'Compound Thoughts' Frege acknowledges that "we really talk figuratively when we transfer the relation of whole and part to thoughts; yet the analogy is so ready to hand and so generally appropriate that we are hardly even bothered by the hitches which occur from time to time". Frege (1923), p. 390. It is, however, quite clear that Frege rejects only the spatial connotations of the part/whole idiom. The idea that thoughts are structured is not a mere metaphor for him.

[20] The same holds for the intensions of Carnap and Montague, which are functions from possible worlds to extensions.

[21] Frege (1891), pp. 135–6, Frege (1892), p. 182, Frege (1906a), p. 302, and Frege (1914), pp. 208–9.

in many different ways – which share chemical properties with hydroxyl. For example, the valence of a radical is *determined* by the valences of its constituent elements and by the structure of their configuration, but it is certainly false that the valence of a radical is *built up* from the valences of its constituents.[22] Correspondingly, if we think of meanings as properties of expressions, rather than as complex objects associated with the expressions, then the building principle loses much of its intuitive appeal. And it is easy to think of meanings in this way; all we have to do is to transform claims of the form 'expression *e* is associated with its meaning *m*' into 'expression *e* has the property of having *m* as its meaning'.

I suspect that the misunderstanding concerning the building principle originates from a certain formulation of the compositionality principle. One might say that the meaning of a complex expression is determined by the meanings of its constituents and the way *they* are combined. Here the antecedent of 'they' is not determinate. If 'they' corefers with 'its constituents', the formulation is equivalent to the principle of compositionality; if 'they' corefers with 'the meanings of its constituents', the formulation is probably best understood as the building principle.

2.3. *The Substitutivity Principle*

Let me restate (4′) from the previous section here as the *substitutivity principle*:[23]

> (S) If two expressions have the same meaning, then substitution of one for the other in a third expression does not change the meaning of the third expression.

This is a crucial thesis, often used in arguing *against* various conceptions of meaning. Take for example the idea that the meaning of a proper name is simply its referent. This implies (together with the *prima facie* innocent assumption that 'Mark Twain' and 'Samuel Clemens' are proper names referring to the same individual) that 'John believes that Mark Twain was American' and 'John believes that Samuel Clemens was American' have the same meaning, which seems clearly false. Or consider the idea that the meaning of a predicate expression is its intension: a function from possible worlds to its extension in that possible world. This implies

[22] The analogy between theories of meaning and theories of valence is from Field (1972), pp. 362–3.

[23] In the philosophical literature the name 'principle of substitutivity' is usually reserved for the thesis that if two singular terms are coreferential then substitution of one for the other in a sentence is truth-preserving. Given the extremely broad understanding of 'meaning' I employed in this paper, this counts as a specific version of (S).

(together with the compelling assumption that 'is the sum of the first 100 natural numbers' and 'is the sum of 5000 and 50' are predicate expressions having the same intension) that 'John believes that 5050 is the sum of the first 100 natural numbers' and 'John believes that 5050 is the sum of 5000 and 50' have the same meaning, which sounds absurd. The arguments obviously use the principle of substitutivity as a tacit premise.

Such arguments may seem simple and persuasive,[24] but the appeal to substitutivity in them cannot be replaced – as it is occasionally assumed – by an appeal to compositionality. As I will argue in this section, the two principles are independent.

It is quite trivial that one cannot get compositionality from substitutivity. Any language that contains *no synonyms* trivially conforms to the principle of substitutivity, but some of these languages are non-compositional. This observation reveals how easy it is to construct examples of languages that violate (C), but not (S). Suppose L is a language where any substitution of synonyms within a third expression leaves the meaning of the third expression unchanged. Let L^+ be an extension of L with some complex expression c and its constituents $e_1, \ldots, e_n$. Let us suppose that c is constructed from $e_1, \ldots, e_n$ using the syntactic rules of L, and that the meaning of c in L^+ is *not* determined by the meanings of $e_1, \ldots, e_n$ in L^+. This means that L^+ violates (C) by stipulation. But if we assume that the expressions $e_1, \ldots, e_n$ have no synonyms in L or among each other, L^+ vacuously conforms to (S).

The only assumption used in this argument is that the characterization I gave of L^+ is coherent. This is not as trivial as it sounds. For one could perhaps argue that the meaning of a complex expression is *trivially* determined by its structure and by the meanings of its constituents if those constituents have no synonyms.[25] But arguing this way would be, I think, a mistake.

As in the argument in Section 2.1, I want to appeal to the ordinary meaning of 'determine' here. Suppose we have a population where everybody has different genes. I take it that it would not be trivial that within this population genes determine eye-color. Notice that if you believe this *is* trivial, you should also believe that within this population, genes determine

[24] It is far from obvious whether such arguments can *prove* that co-referring proper names or coextensional predicates cannot be substituted for one another in belief contexts salva veritate. As the puzzle of Pierre shows, analogous difficulties can be stated without appeal to substitutivity. Cf. Kripke (1979). This may encourage some to bite the bullet and insist that, despite appearances, the substitution pairs in the above paragraph have the same truth-value.

[25] Certainly, if one postulates that (C) says nothing over and above (F), this would be true. But I have already argued that (C) *does* say more than (F).

everything: marital status, precise time of death, whether one will ever see the Taj Mahal, etc. (Why? For there are no two members of the population who are genetically the same, and so *a fortiori*, there are no two members of the population who are genetically the same but differ in whether they will ever see the Taj Mahal.) But I seriously doubt that we would ever say that in a population where all individuals are genetically different from one another, genes trivially determine absolutely everything. So, why say that just because a language happens to be free of synonyms, it is trivially compositional? The question whether the meanings of $e_1, \ldots, e_n$ in L^+ are associated with other expressions as well strikes me as irrelevant in trying to decide whether those meanings (together with the structure of c) determine what c means in L^+. If $e_1, \ldots, e_n$ have no synonyms, we will not be able to use substitution tests in evaluating the claim that the meaning of c is not compositionally determined, but it is not clear to me why this fact would affect the truth or falsity of this claim.

I think the belief that (S) entails (C) is based on an illusion. We are inclined to believe that if in a language substitution of synonyms preserves the meaning of complex expressions, the only conceivable reason for this would be that the language is compositional. The above example shows that the reason might be instead that there are not enough synonyms in the language to detect cases when the meaning of a complex expression is not determined compositionally. If (S) holds for a particular language, this may *suggest* that the language is compositional, but such a consequence cannot be drawn on logical grounds.

The situation is not better with regard to the other direction of the alleged equivalence. Compositionality does not entail substitutivity. Suppose someone suggests that (7) and (8) have the same meaning in English:[26]

(7) Plato was bald.

(8) Baldness was an attribute of Plato.

There are a number or reasonable ways to argue against this position. One can, for example, point out that there are many who would understand (7), but not knowing what attributes are, would be unable to interpret (8). One can point out that (7) is primarily about Plato, whereas (8) is primarily about baldness. Whether these are good arguments is beside the point. (That depends on what notion of meaning one uses.) The following argument is certainly a *bad* one: Assume the compositionality of English,

[26] The example is due to Peter Geach. Cf. Geach (1965), p. 110.

and consider (9). Substituting (8) for (7) in (9) yields (10), which certainly does not mean the same as (9):

(9) The philosopher whose most eminent pupil was Plato was bald.

(10) The philosopher whose most eminent pupil was baldness was an attribute of Plato.

The problem with this argument is, of course, that (7) is not a constituent expression of (9). Compositionality does not guarantee that substituting any expression by a synonym within a complex expression c preserves the meaning of c: the expression that is replaced by another one must at least be a constituent of c. So, perhaps (S) should be modified appropriately, as (S′):

(S′) If e_1 and e_2 have the same meaning and e_1 is a constituent of a complex expression c, then substituting e_2 for e_1 in c does not change the meaning of c.

Unfortunately, such a trivial amendment does not help: (S′) still does not follow from (C). To see why, suppose that someone suggests that the expressions ‘is unrelated’ and ‘is not unrelated’ are synonyms. As an objection, one might point to the fact that the sentences (11) and (12) do not mean the same thing, even though ‘is unrelated’ seems to be a constituent of (11).

(11) Martha is unrelated to everybody.

(12) Martha is not related to everybody.

But such an argument would not be convincing. Intuitively, the difference between the two sentences is a structural one, and it should not be attributed to differences in meaning between ‘is unrelated’ and ‘is not unrelated’. Using a simple formalism, the first sentence is properly interpreted as (11′), while the second as (12′):

(11′) $\forall x(\mathit{person}(x) \rightarrow \neg \mathit{Martha\ is\ related\ to}(x))$

(12′) $\neg\forall x(\mathit{person}(x) \rightarrow \mathit{Martha\ is\ related\ to}(x))$

The problem is that we have no guarantee that substitution of a constituent in a complex expression leaves the structure of the complex expression unaltered. It seems that substitution of ‘is not related’ for ‘is unrelated’ in (11) *does* affect the structure of the sentence. Making certain relatively

uncontroversial assumptions about the syntax of (11) and (12), we can say that the substitution affects whether the quantifier remains in situ or undergoes quantifier raising. And if this is so, the difference in meaning between (11) and (12) cannot prove that their crucial constituents are not synonymous.[27]

Here is another example which illustrates that (S′) does not follow from (C). Imagine that someone suggested that the meaning of a proper name is its referent. Arguing against this proposal, one might point out that since Giorgione was Barbarelli, substitution of synonyms in (13) would yield (14), which – being false – clearly does not mean the same as the true (13).[28]

(13) Giorgione was so-called because of his size.

(14) Barbarelli was so-called because of his size.

Again, this argument seems fishy. (13) means that Giorgione was called 'Giorgione' because of his size. This paraphrase might seem rather removed from the syntactic structure of (13), but it is not. For we can think of the 'so-' in the verb 'so-called' as an indexical element whose interpretation is linked to the actual expression used as the subject of the sentence. The syntactic structure of (13) would then be roughly analogous to the syntactic structure of (13′), where the intended referent of the demonstrative pronoun is the name used in the sentence.[29]

(13′) Giorgione was called that because of his size.

If this is correct, then (13) contains a hidden indexical, and substitution of 'Barbarelli' for 'Giorgione' in the sentence changes the semantic value of this hidden indexical. So the difference in meaning between (13) and (14) fails to show that 'Barbarelli' and 'Giorgione' have different meaning.

The mistaken belief that (C) entails (S) – or (S′) – rests on the idea that all languages resemble simple formal ones. Most formal languages contain only expressions free from any multiplicity of meaning. There are no lexical or syntactic ambiguities, no ellipsis, no contextual parameters. Hence,

[27] Another nice example in the same vein is the following fallacy: Since 'Eve is the mother of Cain' is true and 'Eve's elder son was Cain' is true, therefore 'The mother of Cain's elder son was Cain.' The problem here is that in parsing the expression 'the mother of Cain's elder son', the bracketing [the mother of Cain's][elder son] is syntactically forbidden. This example is from Fine (1984), p. 221.

[28] The example is from Quine (1960), p. 158.

[29] Cf. Crimmins (1992), p. 142.

when one substitutes an expression for another within a larger expression, there is no way this could influence the interpretation of other parts of the larger expression.[30] But the situation is messier in human languages. As the above examples show, substitution may well have an effect on the structure of the larger expression, as well as on the meaning of certain other constituents.[31]

3. Compositionality as Strong Supervenience

We have not yet found a useful elucidation of (C). Given the ordinary understanding of 'determine', (C) excludes the possibility that the meanings of certain complex expressions in a language are fixed by a procedure that involves translation to another language, a possibility that is not excluded by (F). (C) does not entail that the meanings of complex expressions are themselves complex, and hence it is weaker than (B). Finally, (C) seems altogether different from (S): it has nothing to do with the question whether there are enough synonyms in a language, and it does not imply that substitutions of synonyms leave the structure of a complex expression or the meanings of its constituents unchanged.

My aim in this section is to find the appropriate strengthening of (F) that captures what we mean by (C). In order to discover the nature of the connection between (C) and (F), they have to be brought into similar forms. Both of these principles establish some connection between two families of properties. The properties belonging to the first family are properties of *having such-and-such meaning*; I call these meaning properties. The second family contains properties of *having some constituents with such-and-such meanings combined in such-and-such way*; I call these constitution properties. Two expressions are indistinguishable in terms of their meaning properties iff they are synonymous; they are indistinguishable in terms of their constitution properties iff their constituents are synonymous and they have identical syntactic structure. (Note that the definition of

[30] Note, however, that formal in languages containing quantifiers substitutions may lead to bound variable-clashes. In these languages, admissible substitution is defined so that these clashes are always avoided. See fn. 31.

[31] Of course, one might react to these difficulties by *defining* substitution in such a way that a replacement of one expression by another within a third would not count as substitution, unless it is structure-preserving. This would have the unfortunate consequence that the innocent notion of substitution would become theoretically heavy-weight: we no longer have a simple way to decide whether a particular sentence is the result of a substitution within another one. But if we are willing to accept the price, we get a version of (S) which *does* follow from (C). In fact, it would be equivalent to (F).

constitution properties is such that (i) only complex expressions have such properties, and (ii) two different complex expressions can be indistinguishable in terms of their constitution properties even if they contain different (but synonymous) constituents.)

Using this terminology, the two principles formulate what, following G. E. Moore, are called *supervenience-claims* about meaning and constitution-properties. Since the constitution-properties are themselves semantic (they are complex properties that include meaning-properties of lexical items), such supervenience claims do not assert the supervenience of semantic properties on syntactic ones: they merely ensure that certain semantic properties supervene on certain others. I will argue that the difference between (F) and (C) lies in the kind of supervenience relations they postulate.

Consider (F) first. According to this principle, if you take any meaning property – say having the meaning that Julius Caesar was murdered on the ides of March – then for any expression that has that exact meaning, there is a certain constitution property that the expression has, and whatever other expression has that constitution property has the same meaning. Moreover, since this is supposed to be a principle, rather than an accidental true generalization, we have to ascribe some sort of necessity to the claim. The result has the form of what Jaegwon Kim has called *weak supervenience* thesis:[32]

> Necessarily, for any property F in Φ and for any object x, if x has F, then there is a property G in Ψ such that x has G, and if any y has G it has F. In symbols: $\Box(\forall F \in \Phi\ \forall x(Fx \rightarrow \exists G \in \Psi\ (Gx \wedge \forall y(Gy \rightarrow Fy))))$

The standard name for the members of Φ is 'supervenient properties' and Ψ is called the 'supervenience base'. In (F), the supervenient properties are meaning-properties and constitution-properties provide the supervenience base. The values of x and y are complex expressions. So, the function principle can be stated as follows:

> Necessarily, for any meaning property M and any complex expression e, if e has M, then there is a constitution property C such that e has C, and if any complex expression e' has C then e' has M.

Take G. E. Moore's claim that the moral properties supervene on physical ones. According to the weak supervenience scheme, this would mean

[32] Kim (1984), p. 64.

the following: necessarily, whenever you take a moral property, say honesty, then for any person who is honest there is a certain physical property that she has, and whoever else has that exact physical property is also honest. Similarly, the function principle says that necessarily, if you take any meaning property, say having the meaning that Julius Caesar was murdered on the ides of March, then for any complex expression that has that exact meaning, there is a certain constitution property that the expression has, and whatever other complex expression has that constitution property has the same meaning.

The weak supervenience reading of (F) seems to be intuitively correct, but to understand it completely we need an account of the necessity operator in it. Usually, a necessity operator is taken to quantify over possible worlds. This interpretation is out of the question here. *There are* possible worlds at which two complex expressions with the same constitution properties differ in their meaning properties; the actual world is one of them. Why? Because I can actually make up a language where I *stipulate* that 'gray' means gray, 'elephant' means elephant, adjectival constructions have the syntax they have in English, but the expression 'gray elephant' means gray elephant on Sundays, but it means yellow giraffe on other days. Then in this language 'gray elephant' has a single structure and unambiguous constituents, but it is nevertheless ambiguous.[33]

I suggest that the modal force of the function principle is that of quantification over *possible human languages*. Then the function principle says something about which possible languages could be languages *for us*. It asserts that within every possible human language, there is a function from constitution properties to meaning properties. This gives us the following reading for (F):

> For all possible languages L, for any meaning property M and any complex expression e in L, if e has M in L, then there is a constitution property C such that e has C in L, and if any complex expression e in L has C in L then e has M in L.

The principle, understood in this way, implies that a language where there are non-synonymous phrases built from synonymous lexical items using the same syntactic steps is not a possible human language. So, the language mentioned at the end of the previous paragraph is excluded from the family of possible human languages.

The suggested reading for the function principle may seem somewhat counter-intuitive. One might think that the proper way to formulate the principle would have to be relative to particular languages. After all, it

[33] See fn. 4.

seems that whether there are non-synonymous phrases built from synonymous lexical items using the same syntactic steps is a question about English, and has nothing to do with other actual, let alone merely possible human languages. But I think this is not so. The fact (if it is a fact) that there is a such a function in English may have a lot to do with what is going on in other possible human languages. For it may well be the case that the proper *explanation* for why there are no non-synonymous phrases built from synonymous lexical items using the same syntactic steps in English goes like this: The existence of such phrases is logically incompatible with the existence of a function which assigns the meanings of complex expressions to the meanings of their constituents and their structure. However, there is such a function in every possible human language, and since English is one of these, there is one for English as well.

As I already argued in Section 1, the principle of compositionality should be thought of as a general claim about all possible human languages. The argument rests on the observation that our evidence for compositionality – whatever exactly that evidence may be – is not language-specific. The same holds of the function principle, and of many other non-accidental generalizations about particular languages. The reason we believe that the principle holds for English is not that we have observed that there is often a function from the meanings of parts and their mode of combination to the meanings of wholes and so we inductively generalize that this holds for all complex expressions in English. Rather, on the basis of very general considerations, we expect that the principle holds for all possible human languages. (F) is supportable only as a claim of *translinguistic* generality; if true, it is a law of human languages in general.

I argued that the function principle says that meaning properties of complex expressions weakly supervene on their constitution properties. This, however, does not guarantee that the constitution properties of complex expressions *determine* their meaning properties. Kim has argued persuasively that weak supervenience in general fails to capture what we intuitively mean by determination.

According to the thesis that moral properties weakly supervene on physical properties there are no two persons in any possible world who are physically indistinguishable but have different moral properties. However, weak supervenience of moral properties on physical ones is compatible with the supposition that there is a possible world w that is physically indistinguishable from our world, but where some person who is honest in our world is dishonest. But the existence of such a possible world is intuitively incompatible with the claim that one's physical properties determine one's moral properties. The problem is that weak supervenience

of Φ properties on Ψ properties requires only that for every property F in Φ there is a base G in Ψ, but does not say that this G has to be the same in each possible world.

The same point can be raised with regard to the function principle. The principle is not incompatible with the supposition that there is possible human language L which is indistinguishable from some actual language in terms of the constitution properties of its expressions, but where complex expressions would have different meanings. That is, (F) does not exclude Crypto-English as a possible human language. But, as I argued in Section 2.1, the existence of such a possible human language is incompatible with the claim that the meaning of a complex expression is determined by the meanings of its constituents and by its structure.

According to Kim, the strengthening of weak supervenience that we need to approximate the strength of what we mean by determination is the following. He calls this *strong supervenience*:[34]

> Necessarily, for any property F in Φ and for any object x, if x has F, then there is a property G in Ψ such that x has G, and *necessarily* if any y has G it has F. In symbols: $\Box(\forall F \in \Phi\ \forall x(Fx \rightarrow \exists G \in \Psi\ (Gx \wedge \Box\forall y(Gy \rightarrow Fy))))$

(Note that this differs from the weak supervenience thesis only in that it contains a second, embedded occurrence of the necessity operator.)

If moral properties strongly supervene on physical properties, the physical alterego of an honest person cannot be dishonest, even if she happens to inhabit a different possible world. Similarly, if meaning properties strongly supervene on constitution properties, a complex expression constitutionally identical to the English sentence 'Elephants are gray' cannot mean that Julius Caesar was murdered on the ides of March, even if this expression belongs to another possible human language. By applying the strong supervenience scheme to meaning properties and constitution properties and transforming the necessity operators into explicit quantification over possible human languages, we get what I take to be the best paraphrase of the principle of compositionality:

> For all possible languages L, for any meaning property M and any complex expression e in L, if e has M in L, then there is a constitution property C such that e has C in L, and for any possible language L' if any complex expression e' in L' has C in L' then e' has M in L'.

This principle assures that the connection between meaning and constitution properties is the same across all possible languages.

[34] Kim (1984), p. 65.

It is worth noticing that the difference between (F) and (C) according to this interpretation can be stated in another fashion. (F) claims there is no *compositional conflict* among the values of a meaning assignment for a possible human language, i.e., if L is a possible human language and f is its meaning-assignment, then there is no pair of complex expressions c_1 and c_2 in L such that f assigns different values to c_1 and c_2 despite the fact that these expressions are built up in the same way from constituents which receive the same value from f. What (C) says is a straightforward generalization of this claim: there is no compositional conflict among the values of all the meaning assignments of all the possible human languages, i.e., if L and L' are possible human languages and f and f' are their respective meaning-assignments, then there is no pair of complex expressions c_1 in L and c_2 in L' such that the value f assigns to c_1 and the value f' assigns to c_2 are different despite the fact that these expressions are built up in the same way from constituents which receive the same values from f and f', respectively.

If (C) is true and there are no compositional conflicts among the values of all the meaning assignments of all the possible human languages then these meaning-assignments can be combined by union into a single function which compositionally assigns meanings to all the expressions of all the possible human languages. As I understand it, this is exactly what the principle of compositionality says.

Since the meaning-assignments of English and Crypto-English are in compositional conflict, (C) implies that at most one of them can be a possible human language. Is it really plausible to say that if English is compositional Crypto-English is not a possible human language? I believe that it is. For – as I mentioned in Section 1 – it seems plausible that all possible human languages must be learnable by human beings under normal social conditions *as a first language*, and it seems likely that Crypto-English is not learnable in that way. This certainly does not mean that it cannot be learned at all. On the contrary, English-speakers can pick it up almost instantly. But they would learn this language by learning about p, the permutation function which reshuffles the meanings of a few English sentences while keeping the meanings of all other expressions the same. This way of learning Crypto-English presupposes a previous knowledge of English. We have a powerful – and I think entirely justified – intuition that Crypto-English can be learned only as code, as a system that is parasitic on English. I suggest that this is the intuition that underlies our belief in compositionality. And if that is correct, then it is reasonable to interpret the

principle of compositionality as a principle of strong supervenience thesis which implies that Crypto-English is not a possible human language.[35]

Davidson has declared that learnability is a fundamental constraint in the construction of theories of meaning:

> When we regard the meaning of each sentence as a function of a finite number of features of the sentence, we have as insight not only into what there is to be learned; we also understand how an infinite aptitude can be encompassed by finite accomplishments. For suppose that a language lacks this feature; then no matter how many sentences a would be speaker learns to produce and understand, there will remain others whose meanings are not given by the rules already mastered.[36]

Learnability by finite beings requires that the meanings of all expressions be statable in a recursive system. The natural way to try to construct such a theory is to regard certain expressions as semantical primitives and, via semantic rules, to show how the meanings of other expressions are fixed by the meanings of these primitives. It is plausible that for all languages that are learnable *in principle*, we could construct such a theory. However, the acquisition of a great many of these languages may presuppose knowledge of some other language. If Chomsky is right, we cannot learn as a first language a language which violates the principles of universal grammar. And if the principle of compositionally is right, we cannot learn Crypto-English as our first tongue. Any human being who knows Crypto-English must have learned it roughly the way you did: by knowing English and then learning a simple permutation on the sentence meanings.

So I suggest as a useful elucidation of (C) the claim that meaning properties strongly supervene on constitution properties. It is a thesis about languages we speak or could speak as our native tongues, hence it is ultimately a claim about the nature of the human mind.

Before I turn to the conclusion of the paper, I want to address an objection against reading (C) as a strong supervenience thesis.[37] English has a syntactic operation which combines a noun phrase and a verb phrase into a declarative sentence. But there could be another language, say Quenglish, where the same syntactic operation produces the corresponding interrogative sentence. (So, in Quenglish ‘John walks’ would mean what ‘Does John walk?’ means in English.) *Prima facie*, Quenglish seems to be a possible human language that is compositional (assuming, of course, that English

[35] When I presented Crypto-English in Section 2.1, I asked my readers to assume that English is compositional, or to take my use of ‘English’ to refer to a large compositional fragment of English. In the first case it follows immediately that Crypto-English is not a possible human language. In the second, this follows from the additional assumption that rich compositional fragments of English are possible human languages.

[36] Davidson (1965), p. 8.

[37] I tank an anonymous referee of this journal for raising this objection.

is compositional) and also learnable as a first language. But if this is so, Quenglish is a counterexample to (C), as I suggest to interpret it.

There are two possible responses to this challenge: one could either deny that 'John walks' has different meanings in English and in Quenglish, or deny that they have the same structure in these languages. The former option is plausible for *some* conceptions of meaning. The only constraint I put on the notion of meaning was that the meaning of an expression plays a significant role in our understanding of the expression.[38] This does not require that we build the *force* of a sentence into its meaning. If we don't, we might say that the English sentences 'John walks', 'Does John walk?', and 'Walk John!' are synonyms. And then there is no problem with Quenglish being a possible human language.

But what if we want a more inclusive notion of meaning? After all, even if force belongs to pragmatics there is a significant semantic notion of *mood*. If we opt for a notion of sentence-meaning which includes mood, I think we should also opt for a syntactic structure which reveals the mood of the sentence. Mood distinctions are frequently expressed in natural languages by verb infections or specialized lexical items. The accident that in English no declarative marker appears on the surface should not be taken as evidence that English syntax fails to mark declarative mood. After all, the fact that 'I walk' does not carry any visible sign of its tense did not stop syntacticians from assigning a present tense feature to the verb in this sentence.[39] Competent English speakers know that 'John walks' is in declarative mood and I see no reason to deny that this knowledge is syntactic in nature.[40] But then the logical form of 'John walks' must reveal that the sentence is in declarative mood. And if it does, the structure of 'John walks' in Quenglish must be different.[41]

---

[38] Cf. (I) in Section 1.2.

[39] Notice also that it is common to assign focus markers to sentences in logical form, even though in English, focus is only marked by intonation.

[40] Portner (1997) argues persuasively that semantic differences among moods contribute significantly to our understanding of their syntactic distribution.

[41] Jeff King raised a similar objection. Consider a language, say Nenglish, that differs from English only in that the syntactic operation combining a noun phrase and a verb phrase produces a sentence which means the same as its negation does in English. (For example, in Nenglish 'John walks' would mean that John does not walk, even though 'John' and 'walks' mean the same in Nenglish and in English). But if I am right, Nenglish is not a possible human language – a *prima facie* unintuitive result. My response is that on reflection this result may be acceptable: Nenglish is a strange language which probably violates universal principles of human languages. Boolean operators behave surprisingly in Nenglish: 'John walks and John talks' is not logically equivalent to 'John walks and talks' but to 'John walks or talks'. The adjectival detachment inferences are invalid: 'John is a happy person' does not entail 'John is a person'. Finally, there is a worry about negative

4. CONCLUSION

Where does my interpretation of the principle of compositionality leave the debate about the truth of the principle? One might think that the impact is momentous. I suggested that we construe the claim that some particular human language is compositional as a shorthand for the claim that it conforms to the principle of compositionality, and so I am committed to the claim that no human language can be compositional if (C) is false. But if it turned out that Crypto-English is a possible human language, then (C) would turn out to be false and we would have to conclude that no possible human language is compositional.

This sounds rather bold, but it is not nearly as striking as it seems. For, as I said in Section 1, we have only a foggy idea what it is to be a possible human language. Even if it turned out that Crypto-English is in fact learnable as a first language under normal social conditions, that would entail only that it passes *one* test possible human languages must. It may still fail to be a possible human language for reasons unknown.

Understanding (C) in the way I propose will not have much immediate effect on actual semantic theorizing. Its consequences are rather indirect: it brings a certain shift of perspective on the nature of a fundamental principle. If it is accepted that the principle of compositionality is an intricate thesis about the nature of possible human languages, one will be less inclined to think of it as a triviality or as a mere methodological assumption. It will more likely to be taken as a significant, though extremely general empirical assumption.[42]

REFERENCES

Cresswell, M. J.: 1985, *Structured Meanings: The Semantics of Propositional Attitudes*, MIT Press, Cambridge, MA.

Crimmins, M.: 1992, *Talk about Belief*, MIT Press, Cambridge, MA.

polarity items. If we assume that 'John has any friends' is an ungrammatical in Nenglish as it is in English, we cannot hope that this fact can be explained semantically.

[42] I thank the late George Boolos, Richard Cartwright, Molly Diesing, Kai van Fintel, Tamar Szabó Gendler, Michael Glanzberg, Harold Hodes, László Kálmán, Jeff King, Robert Stalnaker, Jason Stanley, Daniel Stoljar, and two anonymous referees for comments and/or discussion on previous drafts of this paper. I am especially grateful to one of the referees whose extensive written comments saved me from errors and prompted essential clarifications. Versions of this material had been presented to audiences at MIT, Cornell, and the University of Rochester. I thank participants in those events for their questions, remarks and objections.

Davidson, D.: 1965, 'Theories of Meaning and Learnable Languages', reprinted in D. Davidson,*Inquiries into Truth and Interpretation*, Clarendon Press, Oxford, 1985, pp. 3–15.

Dekker, P.: 1994. 'Predicate Logic with Anaphora', in M. Harvey and L. Santelmann (eds.), *Proceedings from Semantics and Linguistic Theory IV.* Cornell University Department of Modern Languages and Linguistics, Ithaca, pp. 79–95.

van Eijk, J. and H. Kamp: 1997, 'Representing Discourse in Context', in J. Van Benthem and A. Ter Meulen (eds.), *Handbook of Logic and Language*, Elsevier and Cambridge, Amsterdam, MIT Press, MA, pp. 179–238.

Field, H.: 1972, 'Tarski's Theory of Truth', *Journal of Philosophy* **69**, 347–375.

Fine, K.: 1989, 'The Problem of *De Re* Modality', in J. Almog et al. (eds.), *Themes from Kaplan*, Oxford University Press, Oxford, pp. 197–272.

Frege, G.: 1891, 'On the Law of Inertia', reprinted in B. McGuiness (ed.), *Collected Papers on Mathematics, Logic, and Philosophy*, Blackwell, Oxford, 1984, pp. 123–138.

Frege, G.: 1892, 'On Concept and Object', reprinted in B. McGuiness (ed.), *Collected Papers on Mathematics, Logic, and Philosophy*, Blackwell, Oxford, 1984, pp. 182–194.

Frege, G.: 1906a, 'Foundations of Geometry/II', reprinted in B. McGuiness (ed.), *Collected Papers on Mathematics, Logic, and Philosophy*, Blackwell, Oxford, 1984, pp. 293–340.

Frege, G.: 1906b, 'Introduction to Logic', in Hermes et al. (eds.), *Posthumous Writings*, Chicago University Press, Chicago, 1979, pp. 185–196.

Frege, G.: 1914, 'Logic in Mathematics', in Hermes et al. (eds.), *Posthumous Writings*, Chicago University Press, Chicago, 1979, pp. 201–250.

Frege, G. (1914?) 'Letter to Jourdain', in G. Gabriel et al. (eds.), *Philosophical and Mathematical Correspondence*, Chicago University Press, Chicago, 1980, pp. 78–80.

Frege, G.: 1919, 'Notes for Ludwig Darmstaedter', in Hermes et al. (eds.), *Posthumous Writings*, Chicago University Press, Chicago, 1979, pp. 253–257.

Frege, G.: 1923, 'Compound Thoughts', reprinted in B. McGuiness (ed.), *Collected Papers on Mathematics, Logic, and Philosophy*, Blackwell, Oxford, 1984, pp. 390–406.

Gamut, L. T. F.: 1991, *Logic, Language, and Meaning*, University of Chicago Press, Chicago.

Geach, P. T.: 1965, 'Logical Procedures and the Identity of Expressions', reprinted in *Logic Matters*, University of California Press, Berkeley, CA, 1980, pp. 108–115.

Grandy, R.: 1990, 'Understanding and Compositionality', in J. A. Tomberlin (ed.), *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind*, Ridgeview Publishing Co., Atascadero, CA, pp. 557–572.

Groenendijk J. and M. Stokhoff: 1991, 'Dynamic Predicate Logic', *Linguistics and Philosophy* **14**, 39–100.

Higginbotham, J.: 1986, 'Davidson's Program in Semantics', in E. LePore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Blackwell, Oxford, pp. 29–48.

Hintikka, J.: 1981, 'Theories of Truth and Learnable Languages', in S. Kanger and S. Öhman (eds.), *Philosophy and Grammar*, Reidel, Dordrecht, pp. 37–58.

Jackendoff, R.: 1983, *Semantics and Cognition*, MIT Press, Cambridge, MA.

Janssen, T. M. V.: 1983, *Foundations and Applications of Montague Grammar*, Mathematisch Centrum, Amsterdam.

Janssen, T. M. V.: 1997, 'Compositionality', in J. Van Benthem and A. Ter Meulen (eds.), *Handbook of Logic and Language*, Elsevier, Amsterdam and MIT Press, Cambridge, MA, pp. 417–473.

Kamp, H.: 1981, 'A Theory of Truth and Semantic Representation', in J. Groenendijk and M. Stokhoff (eds.), *Formal Methods in the Study of Natural Language*, Amsterdam Centre, Amsterdam, pp. 277–322.
Kaplan, D.: 1977, 'Demonstratives. An Essay on the Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals', reprinted in J. Perry and H. K. Wettstein (eds.), *Themes from Kaplan*, Oxford University Press, Oxford, 1989, pp. 481–565.
Kazmi, A. and F. J. Pelletier: 1998, 'Is Compositionality Vacuous?' *Linguistics and Philosophy* **21**, 629–633.
Kim, J.: 1984, 'Concepts of Supervenience', reprinted in *Supervenience and Mind*, Cambridge University Press, Cambridge, 1993, pp. 53–78.
Kripke, S.: 1979, 'A Puzzle about Belief', in A. Margalit (ed.), *Meaning and Use*, Reidel, Dordrecht, pp. 239–283.
Montague, R.: 1970, 'Universal Grammar', reprinted in R. Thomason (ed.), *Formal Philosophy*, Yale University Press, New Haven, 1974, pp. 222–246.
Muskens, R.: 1994, 'Compositional Discourse Representation Theory', in P. Dekker and M. Stokhoff (eds.), *Proceedings of the Ninth Amsterdam Colloquium*, ILLC, Amsterdam, pp. 467–486.
Nunberg G., I. A. Sag, and T. Wasow: 1994, 'Idioms', *Language* **70**, 491–538.
Partee, B.: 1984, 'Compositionality', in F. Landman and F. Veltman (eds.), *Varieties of Formal Semantics*, Foris, Dordrecht, pp. 281–312.
Partee, B.: 1988, 'Semantic Facts and Psychological Facts', *Mind and Language* **3**, 43–52.
Pelletier, F. J.: 1994, 'The Principle of Semantic Compositionality', *Topoi* **13**, 11–24.
Portner, P.: 1997, 'The Semantics of Mood, Complementation, and Conversational Force', *Natural Language Semantics* **5**, 167–212.
Quine, W. V. O.: 1960, *Word and Object*, MIT Press, Cambridge, MA.
Salmon, N.: 1989, 'Reference and Information Content: Names and Descriptions', in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic. Vol. 4: Topics in the Philosophy of Language*, Kluwer, Dordrecht, pp. 409–462.
Schiffer, S.: 1987, *The Remnants of Meaning*, MIT Press, Cambridge, MA.
Westerståhl, D., 1998, 'On Mathematical Proofs of the Vacuity of Compositionality', *Linguistics and Philosophy* **21**, 635–643.
Zadrozny, W.: 1994, 'From Compositional to Systematic Semantics', *Linguistics and Philosophy* **17**, 329–342.
Zeevat, H.:1989, 'A Compositional Approach to Discourse Representation Theory', *Linguistics and Philosophy* **12**, 95–131.

The Sage School of Philosophy
218 Goldwin Smith Hall
Cornell University
Ithaca, NY 14853-3201
E-mail: zs15@cornell.edu