Compositionality and Contextuality: The Symbolic and Statistical Theories of Meaning

Yoshihiro Maruyama(✉)

The Hakubi Centre for Advanced Research, Kyoto University, Kyoto, Japan
maruyama@i.h.kyoto-u.ac.jp

Abstract. Compositionality and contextuality give two fundamental principles of linguistic analysis, and yet there is a conflict between them as Burge, Dummett, and others find. Here we aim at elucidating conceptual views underlying their tension in light of both symbolic and statistical paradigms of semantics, arguing, inter alia, that: (i) the conflict is a case of vicious circle analogous to hermeneutic circularity, and may be understood as a tension between symbolic and statistical semantics; (ii) the productivity, systematicity, and learnability of language can be accounted for in accordance with the principle of contextuality as well as compositionality; and (iii) the Chomsky versus Norvig debate on the (symbolic versus statistical) nature of language may be considered a broader manifestation of the tension in the form of the traditional conflict in philosophy between rationalist and empiricist worldviews. We conclude the paper with an outlook for the Kantian synthesis of them, especially the categorical integration of symbolic and statistical AI.

1 Introduction: Two Fregean Principles

Gottlob Frege, one of the founders of symbolic logic and analytic philosophy (especially, philosophy of language), is known for two principles, i.e., the principle of compositionality and the principle of contextuality (even though historical studies [17,25] suggest that both are not strictly rooted in Frege; they then would better be called Fregean principles rather than exactly Frege's). Pelletier [24] summarizes the principle of compositionality as follows: "The Principle of Semantic Compositionality (sometimes called 'Frege's Principle') is the principle that the meaning of a (syntactically complex) whole is a function only of the meanings of its (syntactic) parts together with the manner in which these parts were combined." Another principle of syntactic compositionality may be grounded upon the recursive structure of the syntax of language, which is a character most of the formal and natural languages indeed have (obviously, a text is composed of sentences, which, in turn, are composed of words). Compositionality is considered to be a source of the productivity, systematicity, and learnability of language. Frege [15] says as follows himself:

> I do not believe that we can dispense with the sense of a name in logic; for a proposition must have a sense if it is to be useful. But a proposition consists

G. Bella and P. Bouquet (Eds.): CONTEXT 2019, LNAI 11939, pp. 161–174, 2019.
https://doi.org/10.1007/978-3-030-34974-5_14

> of parts which must somehow contribute to the expression of the sense of the proposition, so they themselves must somehow have a sense. Take the proposition 'Etna is higher than Vesuvius.' This contains the name 'Etna', which occurs also in other propositions, e.g., in the proposition 'Etna is in Sicily'. The possibility of our understanding propositions which we have never heard before rests evidently on this, that we construct the sense of a proposition out of parts that correspond to the words. If we find the same word in two propositions, e.g., 'Etna', then we also recognize something common to the corresponding thoughts, something corresponding to this word. Without this, language in the proper sense would be impossible.

This would suggest that Frege himself grounded the so-called productivity, systematicity, and learnability of language upon the compositional nature of language; the reason why we can understand new sentences we have never heard of before may be accounted for by the compositional nature of language. Concerning these distinctive characteristics of language, Frege [14] also says as follows:

> It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new.

From a historical point of view, however, Janssen [17] argues that Frege did not really endorse the principle of compositionality, but rather endorsed the principle of contextuality (in contrast, Pelletier [25] argues that Frege adopted neither of them). Concerning contextuality, Frege [13] says as follows: "Never ask for the meaning of a word in isolation, but only in the context of a sentence". Wittgenstein [28] also asserts: "Only the proposition has sense; only in the context of a proposition has a name meaning." Notice that this comes from Wittgenstein's early philosophy in *Tractatus*. His later thesis of meaning as use also has some contextualist flavor; our everyday use of language is highly contextual in practice, for example, in that just the same expression can mean different things in different situations or contexts. Wittgenstein's thesis of meaning as use actually inspired what is called statistical distributional semantics, which is probably the most successful contextual semantics in artificial intelligence, especially natural language processing and information retrieval. Yet Chomsky harshly criticizes the nature of contextual statistical semantics in favor of compositional symbolic semantics. Norvig, Google's research director, counters Chomsky in light of the later Wittgensteinian, contextual nature of language. In the following, we fist elucidate the rôles of compositionality and contextuality in our understanding of language, articulating a tension between them, and then shed light on the Chomsky versus Norvig debate from a broader, conceptual perspective.

2 Compositionality Versus Contextuality

Is the nature of meaning in language contextual or compositional? Let us summarize the basic tenets of contextuality and compositionality principles as follows:

- The principle of compositionality: the meaning of a whole (expression) is a function of, and completely determined by, the meaning of its parts (and the syntactical way they are combined together). This is basically atomism; the meaning of atomic expressions recursively generate the meaning of more complex expressions. Compositionality is considered to be a source of the productivity, systematicity, and learnability of language; thanks to compositionality, we can systematically create new expressions.
- The principle of contextuality: the meaning of a word (or more complex expression) is a function of, and can only be determined within, contexts; the meaning of parts depends upon larger wholes surrounding them. This is basically holism about meaning; Quine's holism may be considered a strong form of contextualism. In atomism, parts are prior to wholes, which are secondary; in holism, parts only exist as parts of wholes, which are primary.

The principle of contextuality is often seen as being in conflict with the principle of compositionality (if so Frege cannot endorse both). If meaning is compositional, the meaning of a whole must be determined with reference to the meaning of its parts only, i.e., without reference to anything larger, such as contexts. In contrast, contextuality is a holistic principle. Holism says that a whole cannot be reduced to the mere combination of its parts, whereas the central tenet of compositionality is that this is indeed possible, i.e., the meaning of a whole is composed of that of its parts. Burge [3] is one of the earliest commentators who was clearly aware of the tension in the Fregean philosophy of language:

> It is worth noting that Frege's reasoning here is prima facie *incompatible* with the idea that the notion of the denotation of a term has no other content than that provided by an analysis of the contribution of the term in fixing the denotation (or truth value) of a sentence. The argument presupposes [...] that the notion of term-denotation is more familiar than that of sentence denotation [...]

Dummett [11] also says as follows:

> It was meant to epitomize the way I hoped to reconcile that principle, taken as one relating to sense, with the thesis that the sense of a sentence is built up out of the senses of the words. This is a difficulty which faces most readers of Frege [...] The thesis that a thought is compounded out of parts comes into apparent conflict, not only with the context principle, but also with the priority thesis [...]

Note that Dummett here regards the compositionality of thought as deriving from that of sense (cf. the language of thought hypothesis). The essence of the tension between compositionality and contextuality may be understood in the following manner. The point is whether wholes have to refer to parts or parts have to refer to wholes in oder to determine meaning. Suppose that the two principles are both indispensable in meaning determination. Then, in order to determine meaning, wholes refer to parts, and parts refer to wholes (and mutual reference continues ad infinitum). There is a vicious circle here, and this is essentially an

analogue of what is called hermeneutic circularity in the continental tradition of philosophy. We could speculate that, in the analytic tradition, the two principles came to be understood separately in order to keep the theory of meaning immune to vicious circles, and yet in the continental tradition, both of them were taken at face value at the same time, thus leading to the idea of hermeneutic circularity.

Contemporary developments of semantics in artificial intelligence and machine learning allow us to shed new light on the tension between compositionality and contextuality. Formal semantics today are mostly compositional, whether in linguistics, symbolic logic, or the theory of programming languages. In general, giving semantics is understood to be giving a homomorphism from the algebra of grammar to the algebra of meaning while preserving the compositional structure of language, or in terms of category theory, giving a structure-preserving functor from the category of grammar to the category of meaning. In such developments, both syntax and semantics are compositional, and the issue of contextuality in language is only given a marginal place, or considered to be within the realm of pragmatics rather than proper semantics. Nonetheless, what is dominant in artificial intelligence and machine learning is contextual semantics, called distributional semantics, a conceptual (not strictly historical) origin of which is in Wittgenstein's later philosophy, which puts a strong emphasis on the contextual nature of language in practical use. The Wittgenstein's earlier conception of meaning as correspondence with reality has led to developments of logical semantics. The Wittgenstein's later conception of meaning as use (or meaning in the context of use) has led to developments of statistical semantics today. The relevance of Wittgenstein's later philosophy in contextual statistical semantics in artificial intelligence is much less known than the relevance of his earlier philosophy in compositional symbolic semantics in logic and formal linguistics.

Distributional semantics is based upon what is called the distributional hypothesis, which allow us to generate meaning vectors; the following account builds upon Turney-Pantel [27] (a survey paper cited thousands of times):

- The distributional hypothesis: words that occur in similar contexts have similar meanings. The more contexts, the less possibilities of meaning; this is duality between meaning and context.
- In distributional semantics or the vector space model of meaning, words are represented by vectors, the values of which are determined according to the distributional hypothesis. There are various ways to do this. In the simplest implementation, each value represents how many times the word concerned occur in a given context such as document, sentence, and word co-occurring with it (cf. co-occurrence matrices with each column representing a word and each row a context).
- Similarity between words is given by the inner product of the corresponding meaning vectors, or the relative angle between them. If meaning vectors are parallel, for example, the similarity value is one, which means that they have the same meaning.

Distributional semantics thus gives the linear geometry of meaning. From another angle, we may fix a basis of space, i.e., a set of basic meaning vectors, and then the

weighted sums of them give all meaning vectors, and the weights are given by the distribution of words in different contexts. The distributional hypothesis may be implemented in different ways to compute the weights in practice. Distributional semantics or the VSM (Vector Space Model) based on the distributional hypothesis has made a great success in natural language processing and information retrieval. Turney-Pantel [27], for example, say as follows:

> The success of the VSM for information retrieval has inspired researchers to extend the VSM to other semantic tasks in natural language processing, with impressive results. For instance, Rapp (2003) used a vector-based representation of word meaning to achieve a score of 92.5% on multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL), whereas the average human score was 64.5%.

The conflict between compositionality and contextuality exhibits a tension between symbolic and statistical semantics. Symbolic semantics is based upon compositionalism: the meaning of a whole is the sum of the meanings of its parts (and the way they compose). Distributional semantics is based upon holism: the meaning of parts is determined with reference to wholes. Put another way, it is based upon contextualism: the meaning of the parts is determined with reference to different contexts surrounding them. Frege has established the welll-known distinction between sense and reference. "The evening star" and "the morning star" have the same reference and yet their senses are different. Reference is about what words denote. Symbolic semantics is basically concerned with reference. It is compositional, accounting for meaning while respecting the structure of language such as grammar. Sense is about the mode of presentation, i.e., how words are presented in expressions concerned. Words may have sense without reference (e.g., names of fictional characters). Distributional semantics is more concerned with sense (e.g., meaning vectors for the evening and morning stars are different), accounting for meaning via statistical distribution while ignoring the underlying, generative structure. Let us elaborate the last point in the following. Concerning problems of distributional semantics, Turney-Pantel [27] remark as follows:

> Most of the criticism stems from the fact that term-document and word-context matrices typically ignore word order. In LSA, for instance, a phrase is commonly represented by the sum of the vectors for the individual words in the phrase; hence the phrases house boat and boat house will be represented by the same vector, although they have different meanings.

Note that LSA means Latent Semantic Analysis. Natural language processing, including distributional semantics, is mostly based on so-called "bag of words" models, in which expressions larger than words are seen as multisets of words, thus ignoring word order. It should be emphasized here that such bag-of-words models have achieved great successes in artificial intelligence and natural language processing. Interestingly, Landauer [20] argues that 80% of the meaning of English text is due to word choice and the remaining 20% is due to word

order. Another problem of distributional semantics or contextual semantics in general is that it is difficult to find right representations of non-contextual terms, such as logical connectives. There is thus a trade-off between symbolic logical semantics and statistical semantics. The philosophical tension between compositionality and contextuality manifests as the technical conflict between symbolic logical semantics and statistical distributional semantics. Yet it should not be impossible to overcome the conflict as we shall discuss later.

Davidson [9] also argues that compositionality is indispensable in order to account for the productivity, systematicity, and learnability of language:

> It is conceded by most philosophers of language, and recently by some linguists, that a satisfactory theory of meaning must give an account of how the meanings of sentences depend on the meanings of words. Unless such an account could be supplied for a particular language, it is argued, there would be no explaining the fact that we can learn the language: no explaining the fact that, on mastering a finite vocabulary and a finitely stated set of rules, we are prepared to produce and to understand any of a potential infinitude of sentences. I do not dispute these vague claims, in which I sense more than a kernel of truth.

It is often argued that the productivity, systematicity, and learnability of language could not be accounted for without compositionality (see, e.g., Szabó [26]). Yet we could argue that contextuality is actually able to account for them as well as compositionality. Contextuality could account for the productivity and creativity of language for the following reason. If there were no contextual use of language allowed in our linguistic practice, then the productivity and creativity of language would be lost or weakened. If we could only mean one thing with one expression, our use of language would be more restricted than it actually is. Put another way, we can creatively use just one expression in different situations to mean different things, and this is possible because language is contextual. Contextuality is a source of productivity and creativity. Contextuality never means that the contextual use of language is random, but it rather means that the laws of language are context-dependent, and the contextual laws of language could account for the systematicity of language. And finally our leaning of language is actually highly contextual in practice. Our symbol grounding in the process of language learning is made by associating linguistic expressions with contexts in the world. Above all, the success of machine learning in natural language processing via contextual semantics shows that the principle of contextuality can account for natural language learning. In terms of empirical power so far, systems based upon statistical contextual semantics outperform those based upon logical contextual semantics. It would thus be fair to say that the productivity, systematicity, and learnability of language, in general, can be accounted for by contextuality as well as compositionality, and in certain particular domains, contextuality-based statistical learning systems can even beat compositionality-based symbolic reasoning systems. In the final section we shall think of possible integrations of symbolic AI and statistical AI in natural language processing; before that, we have a look at an interesting debate on the nature of language.

3 Chomsky Versus Norvig on the Nature of Language

There is an insigitful debate between Noam Chomsky, the father of modern linguistics, and Peter Norvig, Google's research director, concerning the nature of language. It is a battle between the symbolic compositional approach and the statistical contextual approach to natural language. From an even broader perspective, it involves a deeper understanding of the nature of science in general, especially the ultimate purpose of the enterprise of science as we shall see below.

Katz [18] interviews Chomsky at MIT, and Chomsky expresses his scepticism about the statistical approach to natural language as follows:

> [I]f you get more and more data, and better and better statistics, you can get a better and better approximation to some immense corpus of text [...] but you learn nothing about the language.

What Chomsky expects to linguistics or science in general is the systematic account of mechanisms or inner workings that gives us a fundamental understanding of phenomena concerned. A statistical approximation to data, by itself, does not give us any understanding. It may be an excellent simulation, but cannot be a scientific explanation, according to Chomsky. Meaning vectors in distributional semantics are constructed based solely upon statistical information about the way linguistic expressions are used in different contexts, in particular how often they co-occur with other expressions. Representing language via meaning vectors, the computer can then solve different problems, actually in a fairly successful and efficient manner. From Chomsky's point of view, however, this is more like semantic engineering rather than semantic science. He nevertheless argues in the interview that statistical data analytics would even be superior to physics in terms of future prediction. Still, what he prefers is a systematic account and understanding rather than purely predictive power.

Gold [16] summarizes the Norvig versus Chomsky debate as follows:

> Recently, Peter Norvig, Google's Director of Research and co-author of the most popular artificial intelligence textbook in the world, wrote a webpage extensively criticizing Noam Chomsky, arguably the most influential linguist in the world. Their disagreement points to a revolution in artificial intelligence that, like many revolutions, threatens to destroy as much as it improves. Chomsky, one of the old guard, wishes for an elegant theory of intelligence and language that looks past human fallibility to try to see simple structure underneath. Norvig, meanwhile, represents the new philosophy: truth by statistics, and simplicity be damned.

To Norvig, what matters is the proper simulation of actual linguistic practice rather than any idealized theoretical analysis. The emphasis on actual linguistic practice is reminiscent of Wittgenstein's later philosophy, which was strongly against any theorizing tendency, and espoused a nuanced, down-to-earth analysis of our linguistic practice in real world situations. Gold [16] seems to be more sympathetic with the Norvig side than with the Chomsky side:

> Norvig is now arguing for an extreme pendulum swing in the other direction, one which is in some ways simpler, and in others, ridiculously more complex. Current speech recognition, machine translation, and other modern AI technologies typically use a model of language that would make Chomskyan linguists cry: for any sequence of words, there is some probability that it will occur in the English language, which we can measure by counting how often its parts appear on the internet. Forget nouns and verbs, rules of conjugation, and so on: deep parsing and logic are the failed techs of yesteryear.

Statistical models do not much take into account the structure of language such as the grammar of language; nonetheless, a purely statistical analysis of co-occurrence of words yields a surprisingly effective model of natural language. Chomsky [7] himself says at an MIT symposium as follows:

> There's a lot work which tries to do sophisticated statistical analysis, you know bayesian and so on and so forth, without any concern for the actual structure of language, as far as I'm aware that only achieves success in a very odd sense of success. There is a notion of success which has developed in computational cognitive science in recent years which I think is novel in the history of science. It interprets success as approximating unanalyzed data.

He admits that statistical models are superior, in terms of predictive power, to other models including symbolic ones. Chomsky [7] further proceeds:

> You would get some kind of prediction of what's likely to happen next, certainly way better than anybody in the physics department could do. Well that's a notion of success which is I think novel, I don't know of anything like it in the history of science. In those terms you get some kind of successes, and if you look at the literature in the field, a lot of these papers are listed as successes. And when you look at them carefully, they're successes in this particular sense, and not the sense that science has ever been interested in. But it does give you ways of approximating unanalyzed data, you know analysis of a corpus and so on and so forth.

To Chomsky, data science may look like a new kind of post-truth science. But science is not just about understanding; it has played crucial rôles in different aspects of human civilization, for example, in wars. If we argue for Norvig, suppose that a war happens between two countries, Chomsky's and Norvig's nations. Which would win the war in the end? Better predictive power would yield better weapons of destruction. The profound understanding Chomsky aims at may be useless to defend his nation. Yet Chomsky could still argue that a deeper understanding of nature results in revolutionary weapons, just as fundamental physics yielded atomic bombs (or as Turing broke the German code) in the WWII.

There is a well-known argument by Chomsky to show that statistical models cannot capture grammaticalness, which he thinks is of purely symbolic nature. To illustrate his point, Chomsky [4] think of the following two sentences, which look similar in form and yet very different in meaning:

– (1) Colorless green ideas sleep furiously.
– (2) Furiously sleep ideas green colorless.

Chomsky [4] then argues as follows:

> It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not grammatical.

Markie [21] explains Chomsky's idea as follows: "Chomsky argues that the experiences available to language learners are far too sparse to account for their knowledge of their language." Chosmky thinks that, without the innate capacity of reason to generate language, we could not even learn language in the first place; this may be called the innate knowledge thesis. Machine learning could be an alleged counterexample to the thesis, but it is not so decisive at the moment. More than fifty years later, Norvig [23] counters Chomsky [4] as follows:

> I'm not sure what he meant by "any of their parts," but certainly every two-word part had occurred, for example: [...] Pereira (2001) showed that such a model, augmented with word categories and trained by expectation maximization on newspaper text, computes that (1) is 200,000 times more probable than (2). I repeated the experiment, using a much cruder model [...] and found that (1) is about 10,000 times more probable.

With respect to this particular point, Norvig can probably win against Chomsky. Yet at the same time, certain deficiencies of present statistical models of language are going to be remedied and overcome with the help of symbolic methods (see, e.g., Coecke et al. [8]). So the complete theory of language and meaning, if any, may require some symbolic methods as well. There is also a huge cost to pay for statistical models, that is, the interpretability of models. Norvig [23] himself says as follows: "I agree that it can be difficult to make sense of a model containing billions of parameters. Certainly a human can't understand such a model by inspecting the values of each parameter individually". Kuhn-Johnson [19] also assert as follows: "Unfortunately, the predictive models that are most powerful are usually the least interpretable." The present situation in statistical machine learning may be comparable to that in quantum mechanics, which comes with strong predictive power and yet with miserably poor interpretation as Feynman-Mermin say "Shut up and calculate!". Kuhn-Johnson [19] also argue:

> If a medical diagnostic is used for such important determinations, patients desire the most accurate prediction possible. As long as complex models are properly validated, it may be improper to use a model that is built for interpretation rather than predictive performance.

There is surely a trade-off between predictability and interpretability. If the primary concern of science lies in empirical prediction, interpretability ought to be

compromised in favor of predictive power. If science is about a rational understanding of nature, on the other hand, interpretability must be maintained at the cost of predictive power. What is at stake here is the very conception of science and its primary aim as an intellectual endeavor of humanity. It should be remarked here that Chomsky and those against statistical data science tend to simplify the actual practice of data science to some extent. They contrast the empirical data-driven approach with the theoretical knowledge-driven approach. Yet data science is not entirely about approximating empirical data. It, for example, attempts to prevent overfitting by introducing domain knowledge in the form of what are called regularization terms or specific priors. It does care about the interpretability of models, and hence the recent trend of interpretable machine learning. The methodology of data science is not purely empiricist but partly rationalist, even if the resulting knowledge is purely empiricist with no explication of underlying mechanisms. A more practical aspect of the cost that complex models have to pay is as follows. Bourguignat [2] says:

> In real organizations, people need dead simple story-telling — Which features are you using? How your algorithms work? What is your strategy? etc. ... If your models are not parsimonious enough, you risk to the audience confidence. Convincing stakeholders is a key driver for success, and people trust what they understand. What's more, at the end of the day, the ultimate goal of the data science work is to put a model in production. If your model is too complicated, this will turn out to be impossible or, at least, very difficult.

Norvig [23] summarizes the points of Chomsky's critique of statistical semantics:

- "Statistical language models have had engineering success, but that is irrelevant to science."
- "Accurately modeling linguistic facts is just butterfly collecting; what matters in science (and specifically linguistics) is the underlying principles."
- "Statistical models are incomprehensible; they provide no insight."

Actually, the metaphor of butterfly collecting as non-science is Chomsky's favorite, and has an older origin. In 1979 Chomsky [6] says:

> You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so.

Chomsky's point may be clarified by having a look at a related case in the history of science, that is, the shift from Ptolemy's predictive model of the heavens to Copernicus's. In terms of predictive power or approximation of data, Copernicus's model did not really outperform Ptolemy's, and Copernicus's nonetheless won the race of science. This is well known among historians of science (see, e.g., Evans [12]). Machine learning papers often claim their methods outperform state-of-the-art methods, in the spirit of what Chomsky calls a new notion of

success in science. If the aim of science is at approximating empirical data, the shift from Ptolemy's to Copernicus's model was unreasonable in light of the very aim of science. In reality, there is some other parameter in the aim of science. Science is concerned with a rational understanding of nature as well as approximating and predicting empirical data. Chomsky aims at the former while Norvig at the latter. The Chosmky versus Norvig debate, therefore, may be seen as a revival of the classic debate between rationalism and empiricism in the context of the contemporary science of language and intelligence. Indeed, Chomsky [5] asserts that he has a "rationalist conception of the nature of language", and contrasts it with an empiricist conception as follows:

> Furthermore, I employed it [...] to support what might fairly be called a rationalist conception of the nature of language [...] In sharp contrast to the rationalist view, we have the classical empiricist assumption that what is innate is (1) certain elementary mechanisms of peripheral processing (a receptor system), and (2) certain analytical machanisms or inductive principles or mechanisms of association.

As the above quote shows, Chosmky is actually aware that what is at stake here is the rationalist versus empiricist conception of language. In the history of philosophy, Kant aimed at reconciling the two camps; Kantian AI (rather than so-called Heideggerian AI) may be a solution to the discrepancy between the two camps. There are recent developments to integrate symbolic AI and statistical AI via different methods; the integrations of symbolic AI and statistical AI may be able to embody Kantian AI, thus allowing us to make predictive power compatible with interpretability. We shall touch upon them in the next section.

Norvig [23] concludes the discussion with the following remark on the contingent nature of language:

> [L]anguages are complex, random, contingent biological processes that are subject to the whims of evolution and cultural change. What constitutes a language is not an eternal ideal form, represented by the settings of a small number of parameters, but rather is the contingent outcome of complex processes. Since they are contingent, it seems they can only be analyzed with probabilistic models.

In his view, language is contingent by nature, and thus statistical models are (not just useful but also) necessary for the analysis of language. By contrast, Chomsky wants to explicate the "eternal ideal form" of language, which is universal rather than contingent. Is the ultimate nature of language contingent or universal? A possible route to do justice to both would be to argue that the surface structure of language is contingent, and yet the core structure of it is universal. Compositional symbolic semantics focuses upon the universal, core structure of language, and contextual distributional semantics upon the contingent, surface structure of language. In this way we could peacefully reconcile the two opposing views.

The tension between Chomsky's and Norvig's views, or between the compositional symbolic approach and the contextual distributional approach, can

be understood in a broader context of philosophy of language. Wittgenstein's early philosophy may be seen as based on compositionalism, and Wittgenstein's late philosophy as based on contextualism. The latter sees meaning as rooted in the "natural history" of linguistic practice (and so dynamic and contingent), whereas the former sees meaning as arising from the picture-theoretical correspondence between language and reality (and thus static and universal). The tension between Chomsky's and Norvig's views is also comparable to the conflict between the realist and antirealist conceptions of meaning in Dummett's terms. Compositional symbolic semantics is usually referential, and presupposes reality outside language, which accommodates denotations to interpret language, and it may be seen as realism. Contextual distributional semantics does not presuppose anything outside language, deriving meaning vectors just from contexts within language. It may thus be seen as antirealism. The autonomy of language as independent of reality is emphasized in Wittgenstein's later philosophy [29]: "The words are not a translation of something else that was there before they were." Meaning comes not from reality, but from the internal structure of language.

Concerning the nature of science rather than language in particular, the tension between Chomsky's and Norvig's views is basically the same as the tension between empiricism and rationalism. As a historical remark, Maxwell had an interesting idea of the relationships between empiricism and rationalism. On one hand, Maxwell [22] says as follows (in his 1850 letter to Lewis Campbell): "the true Logic for this world is the Calculus of Probabilities." Maxwell [22] thinks, perhaps on the basis of the British tradition of empiricism, that human knowledge as a whole rests upon the faculty of sensibility as well as the faculties of reason and understanding, and so it is probabilistic by nature:

> [A]s human knowledge comes by the senses in such a way that the existence of things external is only inferred from the harmonious (not similar) testimony of the different senses, understanding, acting by the laws of right reason, will assign to different truths (or facts, or testimonies, or what shall I call them) different degrees of probability.

Maxwell [22], on the other hand, admits certainty, immutability, and universality, not in empirical experiments, but in things themselves, which can go beyond the world of statistical correlations (though it is not clear whether he thought there was any finitary, human way to access such a world of certainty):

> [O]ur experiments can never give us anything more than statistical information [...] But when we pass from the contemplation of our experiments to that of the molecules themselves, we leave a world of chance and change, and enter a region where everything is certain and immutable.

In such a manner, Maxwell attempted to reconcile the empiricist and rationalist views of science, and it could also allow us to reconcile Chomsky's universalist view and Norvig's probabilist view by regarding the former as concerned with reality per se and the latter as concerned with our surface knowledge of it.

4 Concluding Remarks: Towards the Kantian Synthesis

The principle of compositionality gives the virtually dominant paradigm of semantics in formal linguistics, symbolic logic, and the theory of programming languages. And the principle of contextuality looks marginal from this perspective. Janssen [17] says: "it is not possible to accept both principles at the same time: there is a conflict between them. This is underscored by the fact that the only modern theory which obeys contextuality [...] is proud of being non-compositional and uses this feature to defeat other theories." But what he has in mind is Hintikka's game semantics only. There is another highly successful contextual semantics in artificial intelligence, that is, statistical distributional semantics, which is actually the dominant paradigm of natural language processing, widely used in practical applications. The principle of contextuality gives a mainstream approach to language in artificial intelligence, allowing computer systems to actually learn different features of meaning in natural language. They can even solve some TOEFL problems better than humans. Let us finally discuss whether the two ideas can be combined and integrated into a coherent whole. From a conceptual point of view we can argue that, if language is contextual and contingent in its surface and yet compositional and universal in its core, they are just concerned with different sides of the same coin, and compatible with each other (put another way, neither Chosmky nor Norvig is wrong, and we can do justice to both). There are indeed developments to embody such an idea. One of them relies upon category theory to integrate the compositional and contextual theories of language and meaning (Coecke et al. [8]; one of the earliest integrations). It may be extended to a global paradigm of AI aiming at the categorical integration of symbolic and statistical AI in general, which can, in turn, be regarded as part of an even more global paradigm of categorical unified science qua pluralistic unified science. A similar approach is given by Baroni et al. [1], who adopt "the idea that word meaning can be approximated by the patterns of co-occurrence of words in corpora from statistical semantics, and the idea that compositionality can be captured in terms of a syntax-driven calculus of function application from formal semantics." Domingos et al. [10] also aim to unify logical and statistical AI in a general setting beyond language. All this may be seen as an attempt to achieve the ultimate goal of making empirical performance (Norvig) compatible with systematic understanding (Chomsky). Yet no one (and no machine) can really tell the future. We may possibly end up with the situation that theoretically inclined researchers like Chomsky stick to compositional symbolic semantics, and empirically inclined researchers like Norvig to contextual distributional semantics. In that case it would be justified to conclude that compositionality and contextuality are in true conflict with each other.

References

1. Baroni, M., et al.: Frege in space: a program of compositional distributional semantics. Linguist. Issues Lang. Technol. **9**, 5–110 (2014)

2. Bourguignat, C.: Interpretable VS Powerful Predictive Models: Why We Need Them Both, Medium, 17 September 2014. Accessed on 16 July 2019
3. Burge, T.: Truth, Thought, Reason: Essays on Frege. Clarendon, Oxford (2005)
4. Chomsky, N.: Syntactic Structures. Mouton & Co, Berlin (1957)
5. Chomsky, N.: Recent contributions to the theory of innate ideas. In: Stitch, S. (ed.) Innate Ideas. California University Press, Berkeley (1975)
6. Chomsky, N., Ronat, M.: Language and Responsibility: Based on Conversations with Mitsou Ronat. Pantheon Books, Paris (1979)
7. Chomsky, N.: Keynote Panel: The Golden Age - A Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience, MIT Symposium on Brains, Minds, and Machines. Accessed on 23 June 2019
8. Coecke, B., et al.: Mathematical foundations for a compositional distributional model of meaning. Linguist. Anal. **36**, 345–384 (2010)
9. Davidson, D.: Truth and Meaning. Synthese **17**, 304–323 (1967)
10. Domingos, P., et al.: Unifying logical and statistical AI. In: Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science, pp. 1–11 (2016)
11. Dummett, M.: The Interpretation of Frege's Philosophy. Duckworth, London (1981)
12. Evans, J.: The History and Practice of Ancient Astronomy. Oxford University Press, Oxford (1998)
13. Frege, G.: The Foundations of Arithmetic. Basil Blackwell, Oxford (1953). (originally 1884)
14. Frege, G.: Compound thoughts. Mind **72**, 1–17 (1963). (originally 1923)
15. Frege, G.: The Philosophical and Mathematical Correspondence. University of Chicago Press, Chicago (1980)
16. Gold, K.: Norvig vs. Chomsky and the Fight for the Future of AI, TOR.COM, 21 June 2011. Accessed on 26 June 2019
17. Janssen, T.: Frege, contextuality and compositionality. J. Logic Lang. Inf. **10**, 87–114 (2001)
18. Katz, Y.: Noam Chomsky on Where Artificial Intelligence Went Wrong, The Atlantic, 1 November 2012. Accessed on 23 June 2019
19. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6849-3
20. Landauer, T.: On the computational basis of learning and cognition: arguments from LSA. Psychol. Learn. Motiv. **41**, 43–84 (2002)
21. Markie, P.: Rationalism vs. Empiricism, Stanford Encyclopedia of Philosophy (2017)
22. Maxwell, J.C.: The Scientific Letters and Papers of James Clerk Maxwell: 1846–1862. Cambridge University Press, Cambridge (1990)
23. Norvig, P.: On chomsky and the two cultures of statistical learning. Berechenbarkeit der Welt?, pp. 61–83. Springer, Wiesbaden (2017). https://doi.org/10.1007/978-3-658-12153-2_3
24. Pelletier, F.J.: The principle of semantic compositionality. Topoi **13**, 11–24 (1994)
25. Pelletier, F.J.: Did frege believe frege's principle? J. Logic Lang. Inf. **10**, 87–114 (2001)
26. Szabó, Z.G.: Compositionality, Stanford Encyclopedia of Philosophy (2017)
27. Turney, P., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**, 141–188 (2010)
28. Wittgenstein, L.: Tractatus Logico-Philosophicus. Humanities Press, New York (1961). (originally 1922)
29. Wittgenstein, L.: Zettel. University of California Press, California (1967)