X-bar Theory and the Cantor Set - Theoretical issues in Self-Similarity

Gertjan Postma
(Meertens Instituut)

Abstract
This squib is a theoretical legitimation of the empirical study in self-similarity and quantificational variability (Postma 2020a). We do so by describing and interpreting some parallels between classical X-bar Theory and the Cantor set: they share a fractal geometry and have the same Hausdorff dimension. The splitting of the segment [0;1] of the real numbers $\mathbb{R}$ into the Cantor set part $\mathcal{C}$ (with Lebesgue measure 0) and its complement set $\mathcal{T}$ (with Lebesgue measure 1), provides an overall model of the components of language. It is argued that, if we further split the Cantor set $\mathcal{C}$ into a set of end points $\mathcal{E}$ and a complement set of "internal" points $\mathcal{I}$, an isomorphism can be designed between the triple ($\mathcal{E}$, $\mathcal{I}$, $\mathcal{T}$) and natural language in compositionality, formal semantics (quantification), and lexical semantics. We argue that the structure of the internal part of the Cantor set $\mathcal{I}$, which is self-similar, can serve as a syntactic model in which self-similarity restricts quantificational variability.

1. Introduction

X-bar Theory (Jackendoff 1969, Stowell 1981, Chomsky 1982, Kayne 1994) and the Cantor set (Smith 1874, Cantor 1879) have some properties in common, which we will highlight in this squib. As this paper is written for linguists, we start out exposing the construction of the Cantor set and focus on some well-known properties of it. This part does not contain new insights. Then we describe classical X-bar Theory and interpret it in terms of the Cantor set and show their parallel built. In this part only the reinterpretation and the projection is new. The self-similar part is worked out in a separate empirical study (Postma 2020a).

2. The Cantor Set

The Cantor set was conceived by Henry John Steven Smith in 1874, and some year later thoroughly investigated by Georg Cantor in a sequence of papers from 1879-84.[1] We take a constructivist stand to its nature as it fits best in this introduction. In order to construct the Cantor set, which we will denote by $\mathcal{C}$, one proceeds as follows: start off with the closed set of real numbers from 0 to 1: the closed segment [0;1], split it into three equal thirds, and remove the middle third (open set). We are left with two closed sets [0;1/3] and [2/3;1]. Repeat this process on the two remaining closed segments. Do so ad infinitum. At the n^{th} step in the iteration process, the set C_n is the union of 2^n segments of $(1/3)^n$ in length. The Cantor set $\mathcal{C}$ is the limit

[1] For an accessible introduction, the reader is referred to Chailos (2017). For a historical overview of the origin reasons of constructing this point set, cf. Fleron (1994).

by $n \rightarrow \infty$ of C_n. The various stages of C_0, C_1, … C_7 are visualized in (1).

(1) 

Let us first calculate how much space is removed upon the various steps. In the first iteration step, L_1=1/3 is left out, in the second step, two times 1/9, i.e. L_2 = 2/9, in the third step, four times 1/27, i.e. L_3= 4/27, and so on. This brings us to the series in (2).

(2) $$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \frac{8}{81} + \ldots = L$$

To calculate the sum *L*, one can do the following trick. Multiply both members of the equation with 2/3 (which effectively shifts the sequence one term to the right) and add a new initial term 1/3. The left member of the equation is once again the sequence to be calculated, i.e. (3) holds.

(3) $$L = \frac{2}{3}L + \frac{1}{3}$$

This is easily solved and provides us with $L = 1$. Apparently, $\mathcal{T}$, the union of open sets taken out from [0;1], has length 1, just like the original set [0;1]. In formal terms: the Lebesgue measure of $\mathcal{T}$ is 1. Since 1 – 1 = 0, the Lebesgue measure of the Cantor set $\mathcal{C}$ is 0. So, one might wonder if anything is left after an infinite number of omission steps. One should notice, however, that the end points of the segments are never affected in the subsequent steps, so at least all the end points, e.g. 1/3, 2/3, 1/9, 2/9, 7/9, 8/9, etc. are included in C_3 and in all subsequent steps, and so ad infinitum. Hence, a (countable) infinite number of points (all powers of 1/3), which we denote with $\mathcal{E}$, are included in the Cantor set $\mathcal{C}$, at the least. There are many more points in $\mathcal{C}$, though. An instructive case is the number ¼, discussed here in more detail as it illustrates some interesting features of the Cantor set: self-similarity, its elegant representation in the ternary system, and its relation with quantificational variability of WH, as we will see later.

So, it is worthwhile to consider this case more closely. As we have seen, the end points of the removed open sets are part of the Cantor set. So the point 1/3 and its immediate left is part of the set. In the next step, the end point at point (1/3 – 1/9) is part of the set and its immediate right, and the the next step, the end point at 1/3 – 1/9 + 1/27, etc. The trajectory is

given in red in (4). On every step in the trajectory, the new sum remains in the Cantor set.

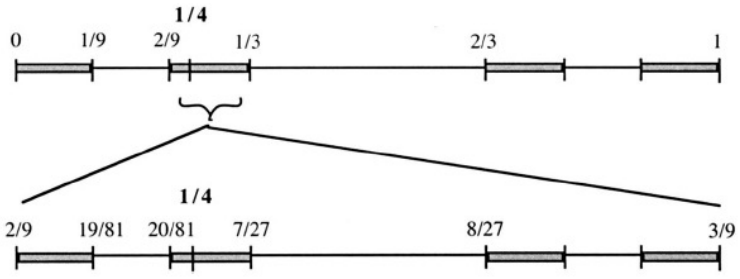
(4a)

(4b)
$$\frac{1}{3} - \frac{1}{3^2} + \frac{1}{3^3} - \frac{1}{3^4} + \ldots = \frac{1}{4}$$

To see that the sequence in (4b) adds up to ¼, assume first it is *n*. Then multiply both members of the equation with a factor –1/3, which effectively shifts the sequence one term to the right. Then add 1/3 to both members of the equation. The left part of the equation is, once again, *n* by hypothesis. This results in the equation in (5), which is easy to solve (under multiplication by 3) and gives the desired result.

(5)
$$n = \frac{1}{3} - \frac{1}{3}n$$
Hence,
$$n = \frac{1}{4}$$

An alternative way to see this result is showing that the path is self-similar upon expansion with a factor 9 or by a reduction with a factor 9. Consider the diagram in (6), taken from Belcastro & Green (2001).



(6)

In this diagram, we observe that the position of the point ¼ with respect to the Cantor set does not change upon scaling by a factor of 9. This is in fact a visual representation of what we have done algebraically above (with scale factor -1/3 and shifting 1/3) in proving the equation in (4b). In this case: scaling with $(-1/3)^2$ and shifting (1/3 – 1/9) = 2/9.

A third way to see that ¼ always remains in the Cantor set during its construction is representing the number in the ternary system. This is an important insight that we will use later in the linguistic representation of WH-items in section 4. In the ternary system, $0 = (0)_3$, $1 = (1)_3$, $2 = (2)_3$, $3 = (10)_3$, $4 = (11)_3$, $5 = (12)_3$, etc. Fractions can, of course, also be converted: $1/3 = 0.333\ldots = (0.1)_3$, $1/9 = 0.111\ldots = (0.01)_3$, etc. In the ternary system, $¼ = (0.0202\ldots)_3$ (cf. Chailos 2017). It can be proven that all numbers that can be cast with only 0s and 2s in the ternary representation belong to the Cantor set. Intuitively, this is understandable because in the first step only representations of the shape 0.0x and 0.2x survive the first deletion step in the Cantor set construction; those with 0.1x, the middle part, are removed. Similarly, only respresentations of the shape 0.00x, 0.02x, 0.20x and 0.22x survive the second step in constructing the Cantor set. Notice that the end point 0.1 can be rewritten as 0.0222… In fact, the 0 and 2 digitals represents the choices in a trajectory of taking a left or right segment route towards the numeral. This can be nicely inspected in the diagram in (4), where one takes an alternating left-right route down the Cantor set: $¼ = (0.0202\ldots)_3$. We will denote the non-end points in C by $\mathcal{I}$, the "internal points". [2]

Before we demonstrate the isomorphism of Cantor set with X-bar Theory, it is useful to have some cardinality considerations. In 1891, Cantor extended cardinality beyond finite sets: also infinite sets have cardinality, which are of several types including the enumarable infinite (HG: *abzählbar*) and uncountable or rather supra-enumerable (HG: *überabzählbar*). The enumerable infinite sets (the set of natural numbers $\mathbb{N}$, the integer set $\mathbb{Z}$, and the set of rational numbers $\mathbb{Q}$) form an equivalence class by the fact that a surjective mapping can be construed from one to the other. These sets have cardinality indicated by $\aleph_0$. The set of real numbers $\mathbb{R}$ has uncountable cardinality (cardinality $\aleph_1$), as does every segment of the real number continuum, e.g. [0;1]. The construction of the Cantor set splits the uncountable domain [0;1] into two subsets: the Cantor set $\mathcal{C}$, as constructivistly defined above, and its complement, which we indicate with $\mathcal{T}$.[3]

We will now calculate the cardinality of $\mathcal{C}$. The cardinality is the same as the cardinality of the entire segment [0;1], i.e. $\aleph_1$. To prove this, we need to construct a surjection from the Cantor set $\mathcal{C}$ onto [0;1]. The standard way to construct such a mapping is to use the finding that $\mathcal{C}$ can be described in the ternary number system by all numerals $(0.x_1x_2x_3\ldots x_i)_3$ for which $x_j \in \{0, 2\}$, because all representations containing 1, are omitted. Now consider the mapping:

---

[2] *Internal* as we use it $\mathcal{C}\backslash\mathcal{E}$, is not identical to *interior*. The Cantor set is without interior.

[3] We call it $\mathcal{T}$ as this set is projects on the terminal symbols in natural language, as we will see.

(7) $\mathcal{C} \rightarrow [0;1]$ by replacing $x_j=0 \rightarrow 0$ and $x_j=2 \rightarrow 1$
for any j in the ternary representation $(0.x_1x_2x_3\ldots x_j \ldots)_3$

… and interpret the result in the binary system. It then covers all binary representations of the segment [0;1]. This shows that $\mathcal{C}$ has the same cardinality as [0;1], i.e. $\aleph_1$.

Let us finally consider the fractal nature of the Cantor set. As we have seen above (cf. 6), the Cantor set is self-similar: it displays scale invariance with subsets of it (Mandelbrot 1983). Such fractal objects have a non-integer dimension. As the Cantor set is a subset of [0;1], and all its elements lay on one line, its dimension must be smaller than 1. Let us first give the generalized definition of *dimension* for fractals, and then apply it to the Cantor set. Let us start out observing that a line has dimension 1, a square dimension 2, and a cube dimension 3. One way to rationalize this assignment, is by considering an object and check by how many objects of a pre-defined set of identical objects are needed to fully cover it. By relating the diameter of the objects to the number that is needed to cover the object, we derive its dimension. A square of diameter 1, for instance, can be covered by 1 square of diameter 1, 4 squares of diameter 1/2, 9 squares of diameter 1/3, etc. Similarly, a cube of diameter 1 can be covered by 1 cube of diameter 1, 8 cubes of diameter 1/2, or 27 cubes of diameter 1/3, etc. In general, the number N_i of objects needed is proportional to d-power of the inverse diameter $(1/r_i)$ of these objects, as given in (8). We call the power *d* the *dimension* of the covered object.

(8)
$$N_i \sim (\frac{1}{r_i})^d \qquad \text{whence,}$$
$$\ln N_i \sim \ln (\frac{1}{r_i})^d = -d \ln r_i \qquad \text{whence,}$$
$$d = -\frac{\ln N_i}{\ln r_i}$$

If we generalize this object-covering strategy to fractal objects by using an iteratively diameter-decreasing set of objects 1, 2, 3,… i, as Hausdorff (1918) proposed, we define the Hausdorff dimension as:

(9)
$$d_H = \lim_{i \rightarrow \infty} \frac{\ln N_i}{\ln r_i}$$

Let us apply this definition to the constructive definition of $\mathcal{C}$. If we decrease the covering objects in every step by $(1/3)^i$, we need a number of 2^i objects to cover the Cantor part (the middle segment does not need to be covered). So the Hausdorff dimension of the Cantor set is

ln2/ln3= 0.6309297536… (cf. Hausdorff 1918:172). For further reference, we list the properties of the Cantor set $\mathcal{C}$ and its complement $\mathcal{T}$ in a table.

(10)

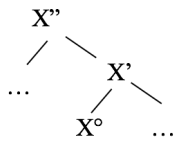
	Cantor Set $\mathcal{C}$	Complement set $\mathcal{T}$ in [0;1]
topology	pointwise discrete	set of continuous segments
iteration	self-similar	terminal
Lebesgue measure	0	1
Hausdorff dimension	0.6309…	1
cardinality	$\aleph_1$	$\aleph_1$, numerable union of $\aleph_1$ sets
subsets	1. $\mathcal{E}$ (end points ($\aleph_0$)) 2. $\mathcal{I}$ (internal points ($\aleph_1$))	

In the next section we compare these finding with the structure of classical X-bar Theory.

3. X-bar Theory

In the study of general phrase structure grammars, it has been a steadily growing insight in the 60's to 80's, that the various rewriting rules of categories repeat the same basic structure (e.g. Stowell (1981). Any phrase is endocentrically composed of a head, a specifier, and a complement. Not all these positions are necessarily visible in an actual sentential structure, but these positions are in principle reserved to be filled or used.[4] This pattern has been generalized to all phrase structures, the functional projections included (e.g. Chomsky 1982 for CP). The traditional name of it is *X-bar Theory*, as the various projections were originally indicated with bars: X°, X' and X". This general structure is given under (11). It has been put on a fundamental setting by Kayne (1994), who derived it form basic linearization assumptions.

(11)

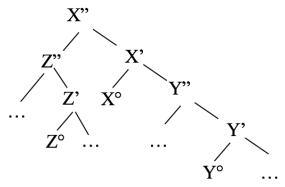


In (11), X° is the *head*, while the higher … is called the *specifier* and the lower … is called the *complement*. The specifier and the complement are feasible of further rewriting through a constituent with similar structure. In other words, the structure is recursive in the specifier (the so-called center-embedding) and recursive in the complement. The head, on the other hand, is a terminal symbol. It is "removed" from the recursion procedure, just like the middle part in the Cantor set. If we proceed this process ad infinitum, we obtain a structure or template on which, idealiter, all possible sentences in all possible languages can be projected. This all according to

[4] For instance by movement, cf. the old concept of *structure preservation*. We leave movement out of consideration in this squib.

classical X-bar Theory.

(12)



This recursive structure spans the entire space of possible sentences. Notice that the head is "removed" from recursive process and spelled out as a terminal. The iteration in the left and right part makes it fully isomorphic with the Cantor Set. We therefore propose to identify the structural part of language with $\mathcal{C}$, and the spell-out component of language with $\mathcal{T}$, the complement set of $\mathcal{C}$, as it is removed from the iteration process. If so, the set of segments in $\mathcal{T}$ realize the subsequent spellouts of Kayne's linearization procedure: the lexical string. We assume that it acquires semantic content through the Lexicon and is identified with the arbitrary, learnable part of language, the Kantian concepts a posteriori. $\mathcal{C}$ itself represent the formal part of language. It is articulated and consists of two subsets: firstly, the set of end points, $\mathcal{E}$, which can be identified with the bracketing of the string. It represents the compositional structure and, hence, the compositional semantics of natural language. Secondly, there is the set of internal points $\mathcal{I}$, e.g. the number ¼, discussed above, of which the interpretation is slightly more difficult. It will be explored in following section.

As to the relation of $\mathcal{E}$ and $\mathcal{T}$, it must be noticed that they are formally separated in the Cantor set and its complement, respectively. On the other hand, they are intertwined in their linearization, provided that we consider the bracketing (the "labelled bracketing" of the old days) part of the linearization: the strict ordering relation defioned on [0;1]. There is another advantage to the conceptual separation of *structure*, represented in the Cantor set, and *content*, represented in its complement: it does no harm to assume that the formal infinite recursive structure is always present, even in those cases in which they do not get a visible spellout. If this idea is correct, every sentence should be modeled as, and approximated by, an infinite recursive object. Especially in a late spellout grammar, such as DM, the formal separation between $\mathcal{C}$ and $\mathcal{T}$ can be taken as a pre-condition of fission-type phenomena, in which a lexeme spells out a complex syntactic structure.

If this formal isomorphism between Cantor set and X-bar structure is correct, it follows that they have the same Hausdorff dimension of 0.603… This is a remarkable result, because one would, at first glance, expect a dimension between 1 and 2. Since grammarians draw

grammatical trees on a sheet, i.e. in a two-dimensional space, its dimension should be ≤ 2. Since it seems richer that a simple string, one expects its dimension to be > 1. From this perspective, a dimension < 1 comes as a surprise. However, as Kayne (1994) has shown, the hierarchical relations and the precedence relations are not independent: linearization is fully determined by hierarchical relations (asymmetric c-command). Kayne's result does not hold vice versa. Our new result, however, shows that one can construct a mapping on a subset of a one-dimensional space: a fractal structure with dimension < 1.

4. The internal set $\mathcal{I}$ and the structure of quantificational meaning

Let us finally reflect on the interpretation of the internal points of the Cantor set, the set $\mathcal{I}$. If our conjecture is correct that the Cantor set $\mathcal{C}$ represents the formal semantic structure of language (such as compositionality) while its complement $\mathcal{T}$ covers the lexicalization, i.e. the arbitrary part of language, we must look into the formal semantic domain to identify the semantic counterpart of $\mathcal{I}$. Let us assume that negation, interrogation, existential quantification do not stem from the outside world, but represent concepts innate to the human species and are imposed upon the world rather than be emergent from the world. They function as tools to interpret the world (Kant 1781). So, how would formal quantificational semantics project on the remaining set $\mathcal{I}$?

Let us go from the known to the unknown. One of the elements that is part of $\mathcal{I}$ is the number ¼, as we have seen above. It represents a self-similar substructure defined over the entire X-bar fractal. In the ternary number representation, it shows up as 0.020202…, a sequence of left- and right-branches ad infinitum. In general, it follows from basic arithmetic facts that all elements $\in \mathcal{I}$ are self-similar in this sense, provided that we limit ourselves to the rational numbers. At this stage, I do not have a clear reason why such a limitation to $\mathbb{Q}$ is defensible. But let us suppose it is. Then all subtrees that do not terminate in a head, i.e. that are not interpreted by the Lexicon, nor are they end points. Hence they must be iterative in some sequence of 0s and 2s, i.e. be a self-similar structure. So we make the conjecture in (13).

(13) Generalized Mapping Hypothesis
- An open variable in the sense of Heim projects on a self-similar subtree.
- Its reading (existential, indefinite, interrogative, etc.) is fully determined by the self-similar pattern

The wording of our conjecture shows that we are inspired by the proposal in Diesing (1992), who ties the readings of indefinites to a syntactic domain, most notably the VP: "existential

closure at VP". Now, the VP sits on a repeated right-branch with respect to all projections in the CP-domain and the TP-domain. Hence, we will rephrase Diesing findings in terms of a self-similar subtree. We characterize such patterns as sequences of 0s (left-branches) and 2s (right-branches), so the VP can be represented as 0.2222. The VP itself is not the variable, but the object or a subject in complement position might be. If so, the self-similar structure characterized by 0.22222 may be present (provided there is absence of scrambling), and a configurational meaning might emerge. An interesting instance of such an open variable is Dutch *wat* 'something', which has next to its existential reading 'something/anything' an interrogative reading 'what?' when it sits in specCP (Postma 1994, Hengeveld, Iatridou & Roelofsen 2019).

(14) a. Jan heeft wat/*WAT gekocht (existential reading / *interrogative reading)
'John has bought something
b. Wat heeft Jan gekocht (*existential reading / interrogative reading)
'what has John bought'

When *wat* 'what' resides in the complement position of VP, it has an existential reading, while the interrogative (echo) reading is ruled out. When it sits in specCP the readings are reversed: only the interrogative reading is available. For an existential reading in specCP one must take recourse to the lexeme *iets* 'something'. Similar data for German (Haspelmath 1997) and many other languages. Curiously, no specific indefinite reading is available to *wat* in the middle field. It seems that only those positions are available which are self-similar: 0.000 and 0.222 respectively. In a subsequent study, we provide more extensive syntactic arguments for this line of research.

Before we go to the conclusions, it is worthwhile to spend a word to the Chomskian discrete infinity of language. The discrete-infinite nature of language is usually illustrated by a recursive structure like *John thinks that Mary thinks that you think* etc. This is a valid but somehow marked example, since the infinite structure has an infinitely long spellout as well. The interesting thing of the parallel with the Cantor set is that it shows another instance of discrete infinity: a *finite* string with an *infinitely* complex internal structure. This instance parallels most other things in nature: the more we look in detail in physics and biology, the complexity we discover. Language is also infinite in this "nanosyntactic" way.

Conclusions

Classical X-bar Theory, as a description of the recursive nature of natural language, is homomorphic with the Cantor set $\mathcal{C} \subset [0;1]$, well studied in mathematics. This means that all

mathematical findings in this branch of research (mainly topology) can be transposed to the structure of language. We generalized the parallel further: we identified the set of open sets $\mathcal{T}$ that were removed from [0;1] during the construction of $\mathcal{C}$, with the lexical string, i.e. the arbitrary, non-recursive part of natural language. The Cantor set constructions splits the formal part and the arbitrary part, that are intertwined in spellout, in a formal way. Furthermore, the Cantor set itself consists of two parts: the set of end points $\mathcal{E}$ whose ternary representation terminates, and the set of internal points $\mathcal{I}$, whose ternary representation does not terminate. The former represents the formal bracketing structure. The latter represents recursive subtrees, which we hypothesized to represent the *syntactic path* of the open variable available for quantificational interpretation.

References

Belcastro, Sarah-Marie & Green Michael (2001). The Cantor Set Contains 1/4? Really? *The College Mathematics Journal* 32, 55-56.

Cantor, Georg (1879-84). Über unendliche, lineare Punktmannigfaltigkeiten. *Mathematische. Annalen* 15 (1879), 1-7; 17 (1880), 355-358; 20 (1882), 113-121; 21 (1883), 51-58, 545-586; 23 (1884), 453-488.

Chailos, George (2017). *Cantor Set – A Naïve Introduction.* Presentation for π-Day, 2017.

Chomsky, Noam (1982). *Some concepts and consequences of the Theory of Government and Binding.* Linguistic Inquiry Monograph 6. MIT Press.

Diesing, Molly (1992). *Indefinites*. Linguistic Inquiry Monograph 20. MIT Press.

Fleron, Julian F. (1994) A Note on the History of the Cantor Set and Cantor Function *Mathematics Magazine* 67(2) 136-140.

Hausdorff, Felix (1918). Dimension und äußeres Maß. *Math. Annalen* 79 (1918), 157-179.

Heim, Irene (1982) *The semantics of definite and indefinite noun phrases.* PhD Dissertation, Univ. of Massachusetts at Amherts.

Hengeveld, Kees, Sabine Iatridou & F. Roelofsen (2019). *Quexistentials I*. Manuscript UvA.

Kant, Immanuel (1781). *Kritik der reinen Vernunft*. Riga.

Kayne, Richard S (1994). *The Antisymmetry of Syntax*. LI Monograph 25. MIT Press.

Mandelbrot, B. (1983). *The Fractal Geometry of Nature.* W. H. Freeman, San Francisco.

Postma, Gertjan (1994). The indefinite reading of WH. *Linguistics in the Netherlands* 1994. 187-198.

Postma, Gertjan (2020b) *Empirical Issues in Self-Similarity and Quantificational Variability.* Ms Meertens Institute.

Schayer, Rhiannon & David Jordan (2003). Rational Points In The Cantor Middle Thirds Set.

Smith, Henry John Steven Smith (1874). On the Integration of Discontinuous Functions. *Proceedings London Mathematical Society*, (1875) 140–153.

Stowell, Tim (1981). *Origins of Phrase Structure.* PhD Dissertation. MIT.