Content and the Internet, Part II

by
Del Jensen and Steve Carter

The Computational Semantic Model of the Lexicon

Synopsis

> *"Certain mystes aver that the real world has been constructed by the human mind, since our ways are governed by the artificial categories into which we place essentially undifferentiated things, things weaker than our words for them."*
> *- Gene Wolfe*

In *Part I* we introduced the idea of a *semantic appliance*: a device which could be positioned, for example, in front of queries directed to large content stores or at the border of a network. The appliance would provide a *semantic map* of the network content (i.e., content interior to the store or border). Ideally, the appliance would scale (levels of abstraction) to provide service across all layers of the Internet. With such a device, people and agents could search for content with *meaning* factored into the search criteria. We mentioned the semantic appliance as one example of what can be done with *a normative computational semantic model for content.*

> . . . like meanings are close together and unlike meanings are far apart.

To be effective, a computational semantic model must establish a *quantitative context* for that which is *essentially qualitative* in nature. Furthermore, the model should preserve what Saussure [1] termed “relational identity” within “the conceptual field.” In other words, we need a computational model wherein *like meanings are close together* and *unlike meanings are far apart*. Meanings of what? In our case we would like to know the meanings of the various linguistic elements (lexical, sentential and discursive) that in synthesis form the content, content which is primarily natural-language in form (e.g., HTML).

In mathematical terms, the semantic mapping should be *continuous* and *one-to-one*. A mapping that is both continuous and one-to-one is called a *homeomorphism*, and it is a concept which figures prominently in the mathematical discipline called *point-set-topology*. Topology uses language and logic in a precise way in order to formalize, refine and extend the human-intuitive notion of *space* into a *useful* body of knowledge. By space, we mean a context for discourse (about *things*). By useful, we mean public confidence that the discipline is *self consistent* and that it have a quality of *universality*. Stated more precisely and less pedantically, "Suppose that a mathematician is confronted by some concrete collection of objects into which he wishes to introduce a topology so that, for example, he may define continuous functions on the collection. There may be many ways to do this, but it is usually convenient to introduce a topology that is as 'strong' as possible in the sense that much is known about the particular topology. For instance, if it were possible to topologize the collection as a compact metric space, our mathematician would probably do so."

[Hocking, John G. and Young, Gail S., Topology, Addison-Wesley 1961]

We previously outlined a twofold representation for *meaning* at the sentential level:

- *A Model for Lexical Semantics* - wherein we embed the lexicon (in the form of meaning postulates) in a complete metric space (**S**, d),
- *A Model for Phrase Semantics* - wherein we identify invariant grammatical elements (via Universal or Transformational Grammar) as clusters of points in the lexical space (**S**, d), and abstract the context to the derived space (,(**S**), h) of compact sets on (**S**, d).

In this paper we further develop the theme of *modeling the lexicon*. We propose a specific mechanism for metrization, and apply the technique to a simple example as a demonstration of effectiveness.

Lexical Semantics

Words! Mere words! How terrible they were! How clear, and vivid, and cruel! One could not escape them. And yet what a subtle magic there was in them! They seemed to be able to give a plastic form to formless things, and to have a music of their own as sweet as that of viol or of lute. Mere words! Was there anything so real as words?

- Oscar Wilde

Our aim is not so lofty as that of the theorist, and we should keep this fact in mind. We are not trying to explain the origin of human language, neither are we expounding a theory on the nature of human thought. All we want is a framework where we can relate lexical elements to mathematical objects. The relation should be such that lexical items that are "close" in meaning correspond to objects that are "close" in a mathematically computable sense, and lexical items that are "different" correspond to objects that are "far" in a computable sense. With such a model we could manipulate lexical elements via operators in the transform space. But before we proceed to develop our model, we should take a quick look at the current status of lexical semantics.

Eco [2] marshals several arguments against a dictionary model of the lexicon. He examines several hierarchical models and discards them all, demonstrating that no *single tree* can encompass the nature of meaning. He takes the view that only an encyclopedic model can prevail, in the form of a labyrinth or "net!"

By contrast, Chomsky writes, "The problem [for signs] is not one of vagueness; rather of hopeless underspecification." [3] He goes on to say, "In general, a linguistic expression provides a complex perspective …, the conclusion only becomes clearer as we move from the simplest case – proper names and common nouns – to words with inherent relational structures and more complex constructions." [4] All this leads up to his (not surprising) assertion that meaning can not stand aloof from syntax [5]. But Chomsky acknowledges that organisms have a "cognitive space." [6]

Bilgrami supports Chomsky's position, pointing out that Chomsky has been criticized by those "who have consistently said that he has unnecessarily restricted scientific and formal theory to certain aspects of syntax and formal semantics leaving the lexical aspect of semantics ('meaning') out in the cold." He goes on to say, "[Chomsky argues] that the lexical aspects of language are to be thought of as bringing in an agent's perspective on things in the world." [7]

[We] adopt a *directed set* as the lexical model.

Our response to Eco is to advocate his position, and adopt a *directed set* as the lexical model. The topological structure of our model is derived from the natural tension between two categorical imperatives: *identity* and *difference*. In this, we follow the lead of Sassure [8], who pointed out the inevitably preeminent roles of identity and difference given that the "conceptual field" is a continuum. With this approach, Chomsky's objection of underspecification seems to be a restatement of Saussure's principle of the "arbitrary" nature of the sign [9]. Chomsky's position on whether *meaning* has any property of continuum is unclear to us. Indeed, he directly attacks the notion that "… language use and acquisition … are somehow 'based on meaning,' the idea being that semantics is soft and mushy … " [10]. We agree with him in the sense that with contextual and syntactic cues resolved, the semantic characteristics of a lexical element map unambiguously to a region of the conceptual field; i.e., *arbitrary* does not necessarily equal *soft and mushy*.
In Chomsky*ese*, we propose a model which maps a semantic aspect (of the D-Structure) across the S-Structure and LF level to a cognitive space in the Mathematical Faculty. As we shall see, our *cognitive space* is simply a reified *conceptual field*.

. . . our *cognitive space* is simply a reified *conceptual field*.

Because the model that we propose is essentially topological in its nature, our response to Bilgrami's interpretation of Chomsky is that a topology <u>is</u> an abstraction of a perspective on things. Of course meaning (of words) cannot always be resolved indepedently of syntax. In part this is due to symbolic overload. For example, consider the two phrases: "we heard the sonic boom," and "the shadow of the sound boom ruined the entire scene." The phrases refer to unrelated meaning postulates for "boom." Even more difficult to resolve are issues of metonymy and metaphor.   But this is an aspect of computational linguistics that is well developed, and we are neither ambitious nor presumptuous enough to propose yet another syntactic model; rather, we proceed on the assumption that we will use one of the many excellent available implementations.

At any rate, in our attempt to project the lexical aspect of our "language faculty" onto our "mathematical faculty" [11] we can be accused of nothing worse than "having no more than the morals of engineers." [12] Our task is to demonstrate that enough "word stuff" survives the projection to be useful to the empirically minded.

Of course, when embarking on any kind of grand engineering endeavor we must be mindful of the danger of hubris …

A Computational Semantic Model

… and they said, "Come, let us build ourselves a city, with a tower …"
- Genesis, 11:4

In our model, hyponymy - the axis of *identity as propositional entailment* – defines the topological essence of the lexicon, subject to the paradigmatic constraints of separability. In other words, we substitute Saussure's notion of *identity* with the "yin" of *openness,* and his concept of *difference* with the "yang" of the *separability axioms*. Guided by these two principles, we build a qualitative model of the lexicon in the form of a directed set. We then exploit the classic metrization theorems of point-set topology to map our qualitative model into a subspace of the Hilbert coordinate space. The Hilbert subspace, with its associated metric and inner product, is our proposed *computational semantic model of the lexicon*, (**S**, d). This is the program.

Defining a Topology for **S**

We use hyponymy to define our topology. For example, the set s_1 = {*men*, *women*, *boys*, *girls*} is superordinated by the set corresponding to *human beings* which at least includes {*men*, *women*, *boys*, *girls*, *youth*, *toddlers*, *infants*}, in the sense that every element of s_1 *entails* a set of attributes that is common to every element of *human beings*. Since s_1 is contained in *human being* as an analytically true proposition (*human being* as *meaning postulate*), we say that s_1 is an element of the basis for the topology, and require that it too be a meaning postulate. An alternative formal symbol for s_1 might be *humans that are (typically) continent*.

While the set {*husbands*, *bachelor U.S. presidents*} is contained in *men*, and so is an element of the basis, the set {*husbands*, *U.S. presidents*} does not qualify in relation to *men* as a basis element, since it is not analytically true that *U.S. presidents are men*. However, *people* does contain {*husbands*, *U.S. presidents*}, making the latter set necessarily an element of the basis. Of course, this structure incorporates the authors' personal "perspective on things." That's all right; if someone else expresses a topology wherein *men* does contain {*husbands*, *U.S. presidents*}, there will be a kind of "impedence" mismatch of elements in this neighborhood of the lexicon (which can be resolved if we compare the two topologies). But this simply parallels the "sloppiness" of everyday speech between people.

We use *identity* (as hyponymy) to generate a topological basis; but how do we factor *difference* into the model? By way of example, consider the intersection of the category *adult* with *human beings*. This is certainly a valid meaning postulate. What does this set "look like?" We know that it contains *men* and *women*, but it is not clear that it is <u>equal</u> to the union of *men* and *women*. If we further stipulate that the intersection of *men* and *women* is the empty set, then the inclusion chain {*men*, *women*} $\subseteq$ *adult* $\cap$ *human beings* starts to have some interesting structure.

We are at a point where a concrete example of a (very restricted) lexicon is in order. Consider the qualitative model shown in figure 1.

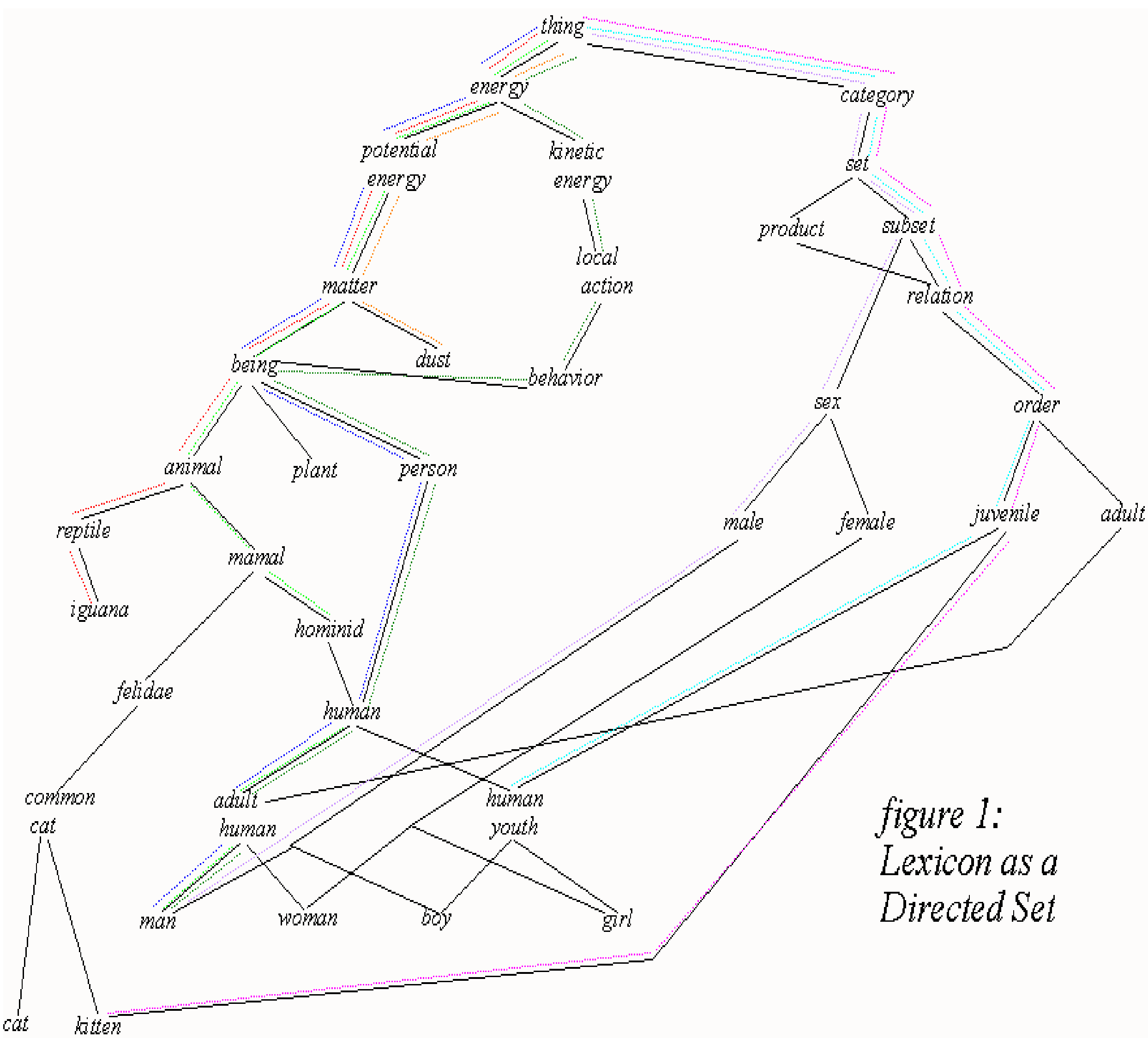

figure 1: Lexicon as a Directed Set

Some observations about the nature of figure 1:

- First, we claim that the model is a *topological space*.

- Second, note that *the model is not a tree*. In fact, it is an example of a *directed set* [13].

- Third, observe that the relationships expressed by the links are indeed relations of hyponymy.

- Fourth, note particularly – but without any loss of generality - that *man* maps to both *energy* and *category* (via composite mappings) which in turn both map to *thing*; i.e., the (composite) relations are multiple valued and induce a partial ordering. These multiple mappings are natural to the meaning of things and critical to semantic characterization.

- Finally, note that *thing* is *maximal*; indeed, *thing* is the *greatest* element [14] of *any* quantization of the lexical semantic field (subject to the premises of our model).

We urge the reader to verify the above points. In order to metrize the set we employ a common tool of abstraction known as *reification*, and understanding will be greatly enhanced if one first takes time to gain a bit of familiarity with the model.

> . . . we employ a common tool of abstraction known as *reification*. . . .

Metrizing **S**

The colored paths shown in the figure suggest that the relation between any node of the model and the maximal element *thing* can be expressed as any one of a set of *composite* functions; one function for each chain from the minimal node μ to *thing* (the nth predecessor of μ along the chain):

$$f: \mu \Rightarrow thing = f_1 \circ f_2 \circ f_3 \circ \cdots \circ f_n$$

(see figure 2) where $n+1$ is the cardinality of the chain, and

$$f_j: (n\text{-}j)\text{th predecessor of } \mu \Rightarrow (n+1\text{-}j)\text{th predecessor of } \mu,\ 1 \le j \le n.$$

Consider the set of all such functions for all minimal nodes. Choose a countable subset $\{f_k\}$ of functions from the set. For each f_k we construct a function g_k: S $\Rightarrow \mathbb{R}^1$ as follows. For $s \in$ S, s is in relation (under hyponymy) to *thing*. Therefore, s is in relation to at least one predecessor of μ, the minimal element of the (unique) chain associated with f_k. Then there is a predecessor of smallest index (of μ), say the mth, that is in relation to s. We define $g_k(s) = (n-m) / n$.

Finally we can define the vector valued function φ: S $\Rightarrow \mathbb{R}^k$ relative to the indexed set of scalar functions $\{ g_1, g_2, g_3, \ldots, g_k\}$; $\varphi(s) = \langle g_1(s), g_2(s), g_3(s), \ldots, g_k(s)\rangle$.

$$\varphi(s) = \langle g_1(s),\ g_2(s),\ g_3(s),\ \ldots g_k(s)\rangle$$

The functions g_k are analogous to step functions, and in the limit (of refinements of the topology) the functions are continuous. Continuous functions preserve local topology; i.e., "close things" in S map to "close things" in |^k, and "far things" in S tend to map to "far things" in |^k.

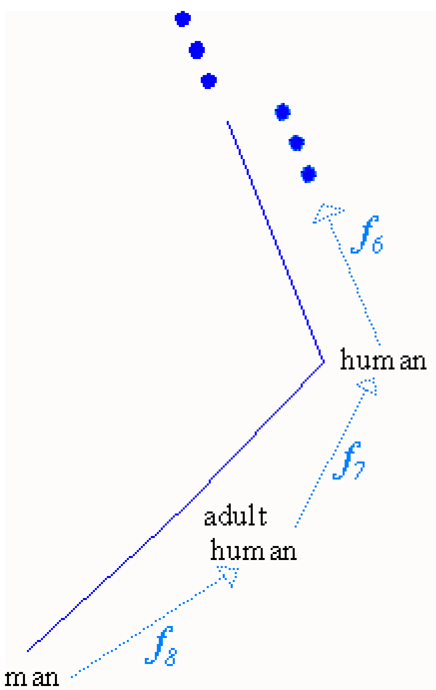

***figure 2*:**
The blue chain,
$\mathbf{f} = f_1 \circ \cdots \circ f_6 \circ f_7 \circ f_8$
with minimal element μ = man.

Example Results

Some items from the example of figure 1:

		g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
		♦	♦	♦	♦	♦	♦	♦	♦
φ(*boy*)	⇒	⟨3/4,	5/7,	4/5,	3/4,	7/9,	5/6,	1,	6/7⟩
φ(*dust*)	⇒	⟨3/8,	3/7,	3/10,	1,	1/9,	0,	0,	0⟩
φ(*iguana*)	⇒	⟨1/2,	1,	1/2,	3/4,	5/9,	0,	0,	0⟩
φ(*woman*)	⇒	⟨7/8,	5/7,	9/10,	3/4,	8/9,	2/3,	5/7,	5/7⟩
φ(*man*)	⇒	⟨ 1,	5/7,	1,	3/4,	1,	1,	5/7,	5/7⟩

<u>Distance</u> <u>Angle Subtended</u>

|| boy – dust ||	~ 1.85	~52°
|| boy – iguana||	~ 1.65	~46°
|| boy – woman||	~ .41	~10°
|| dust – iguana||	~ .80	~30°
|| dust – woman||	~ 1.68	~48°
|| iguana – woman||	~ 1.40	~39°
|| man – woman||	~ .39	~07°

Some Comparisons:

- Boy is closer to Iguana than to Dust.
- Boy is closer to Iguana than Woman is to Dust.
- Boy is much closer to Woman than to Iguana or Dust.
- Dust is further from Iguana than Boy to Woman or Man to Woman.
- Woman is closer to Iguana than to Dust.
- Woman is closer to Iguana than Boy is to Dust.
- Man is closer to Woman than Boy to Woman.

All other tests done to date yield similar results. The technique works consistently well.

How It (really) Works

As described above, construction of the φ transform is (very nearly) an algorithm. In effect, we have described a *recipe* for metrizing a lexicon – or for that matter, metrizing anything that can be modeled as a directed set – but we haven't addressed the issue of *why* it works. In other words, *what's really going on here?* To answer this question, we must look to the underlying mathematical principles.

First of all, what is the nature of **S**? In *Content and the Internet, part one*, we suggested that a propositional model of the lexicon has found favor with many linguists. For example, the lexical element *automobile* might be modeled as:

{automobile: *is a* machine,
is a vehicle,
has engine,
has brakes,
•
•
•
}

In principle, there might be infinitely many such properties, though practically speaking we might restrict the cardinality to ✐$_0$ (countably infinite) in order to ensure that the properties are addressable. If we were disposed to do so, we might require that there be only finitely many properties associated with a lexical element. However, we see no compelling reason to require finiteness.

At any rate, we can see that *automobile* is simply an element of the power set of **P**, the set of all propositions; i.e., it is an element of the set of all subsets of **P**. We denote the power set as –*(***P***)*. Note that the first two properties of our *automobile* example express *is a* relationships. By *is a*, we mean entailment. By *entailment* we mean that, were we were to intersect the properties of every element of –*(***P***)* that we call, for example, *machine*, then the intersection would contain a subset of properties common to anything (in –*(***P***)*) that we have, do, will or would have called *machine*. Our reliance on the existence of a "least" common subset of properties to define entailment has a hint of *well ordering* about it [15]; and indeed it is true that we rely on the *axiom of choice* to define entailment!

For the moment, let us restrict the notion of *meaning postulate* to that of entailment. To be sure, there are many other kinds of rules - rules that are, arguably, more interesting than entailment - and we will address some of these rules at a later time. Let **B** = $\{b_\alpha\}$ be the set of elements of –(–*(***P***))* that correspond to *good* meaning postulates; e.g., $b_m \in$ **B** is the set of all elements of –*(***P***)* that entail *machine*. By good, we mean *complete* and *consistent*. By complete, we mean non-exclusionary of objects that *should* entail (some concept). By consistent, we mean exclusionary of objects that *should not* entail (any concept). We understand *should/should-not* to be negotiated among and between the community (of language users) and its individuals.

Note that if the intersection of b_β and b_γ is non-empty, then $b_\beta \cap b_\gamma$ is a "good" meaning postulate, and so must be in **B**. We define the set **S** = $\cup\, b_\alpha$ to be the lexicon. A point of **S** is an element of –*(***P***)* that entails at least one meaning postulate.

We constructed **B** deliberately to be the basis of a topology τ for **S** [16]. In other words, an open set in **S** is defined to be the union of elements of **B**. This is what we mean when we say that *we use hyponymy to define the topology of the lexicon*.

The separability properties of **S** are reflected in the *Genus / Species* relationships of the unfolding inclusion chains. We adopt the T_0 – T_4 trennungsaxioms, and turn our attention to the set of bounded continuous real valued functions on **S**. In the words of Hocking and Young [17]:

> " …, for any space, the collection of bounded continuous real valued functions on the space can be made into a metric space whether or not the original space is metric. This might suggest that questions about a given space with some weird topology perhaps can be answered by investigating a rather nice function space. *An obvious requirement here would be to have enough continuous functions to be able to distinguish between the points of the given space*." – italics added

To this end, we cite Urysohn's lemma.

- **Theorem**. If **S** is a normal space and *A* and *B* are two disjoint closed subsets of **S**, then there is a real-valued continuous function *g*: $\mathbf{S} \Rightarrow \mathbf{I}^1$ of **S** into the unit interval $\mathbf{I}^1$ such that *g(A)* = 0 and *g(B)* = 1.

Our use of *g* to denote the function was not accidental; it should evoke the scalar coordinate functions $\{ g_1, g_2, g_3, \dots , g_k\}$ of the previous section. A proof of the lemma can be found in almost any elementary general topology book; in particular, see Hocking and Young [18].

We are very nearly where we want to be! Before we can invoke a final theorem of Urysohn's and complete the metrization of **S**, we must introduce the notion of a Hilbert coordinate space.

Consider the set **H** of all sequences $\gamma = \{\gamma_1, \gamma_2, \gamma_3, \dots\}$ of real numbers such that $\Sigma^\infty \gamma_i{}^2$ converges. We define the metric

$$d(\gamma, \chi) = \Big[\sum_{I=1}^{\infty} (\gamma_i - \chi_i)^2 \Big]^{1/2}.$$

on the set **H**, and denote the Hilbert coordinate space (**H**, d).

If we look at the sequence $\{\gamma_1, \gamma_2, \gamma_3, \dots\}$ as a *vector*, we can think of Hilbert space as a kind of "super" Euclidean space. Defining vector addition and scalar multiplication in the usual way, it is no great feat to show that the resultant vector is in **H**. Note that the standard inner product works just fine.

We are looking for a metric space equivalent to our topolgical space (**S**, τ), and Urysohn's lemma should be a strong hint to the reader that perhaps we should be looking at (**H**, d).

- **Theorem**. Every completely separable normal space **S** is homeomorphic to a subspace of Hilbert's coordinate space.

We prove the theorem by actually constructing the homeomorphism.

Proof: Let $B_1, B_2, \dots, B_n, \dots$ be a countable basis for **S**. In view of Theorem 2-4, there are pairs B_i, B_j, such that $[B_i]$ is contained in B_j; in fact, each point of **S** lies in infinitely many such pairs, or is itself an open set. However, there are at most a countable number of pairs for each point of **S**. For each such pair B_i, B_j, Urysohn's lemma provides a function g_n of **S** into $\mathbf{I}^1$ with the property that $g_n([B_i]) = 0$ and $g_n(S-B_j) = 1$. (If the point p forms an open set, then we take $g_n = 0$ for large *n*.) Letting **H** denote the Hilbert coordinate space, we define the (vector valued) mapping ϑ of **S** into **H** by setting

$$\vartheta(s) = \{g_1(s), g_2(s)/2, g_3(s)/3, \dots, g_n(s)/n, \dots\}$$

for each point *s* in **S**. It remains to prove that the function ϑ so defined is continuous, one-to-one, and open. [19]

The original proof (in its entirety) is available in the literature [20]. When ϑ is applied to a lexicon with the entailment topology, we call it the *Bohm transform*. Clearly, the finite-dimensional transform φ is an approximation of the Bohm transform, mapping the explicate order of the lexicon to a (shallow) implicate order in $\mathbb{R}^k$.

> . . . the . . . transform φ is an approximation of the Bohm transform, . . .

Here, reader, we stop and rest for a time. Were you to leave us now in our journey toward that shadow-land called meaning, we would not blame you. It is no easy road.

References

1. Culler, Jonathan, *Ferdinand de Saussure*, 39-45, Cornell University Press (1986).
2. Eco, Umberto, *Semiotics and the Philosophy of Language*, 46-86, Indiana University Press (1986).
3. Chomsky, Noam, *Language and Thought*, 19, Moyer Bell (1994).
4. ibid, p. 23.
5. ibid, pp. 24, 26-27.
6. ibid, p. 44.
7. ibid, p. 61.
8. see [1], pp. 45-57.
9. see [3], p. 19.
10. [3], p. 26.
11. Chomsky, Noam, *The Minimalist Program*, 13-64, The MIT Press (1995).
12. Wilks, Yorick, "Reference and its Role in Computational Models of Mental Representations," in Eco, U. and Violi, M.S.P. (editors), *Meaning and Mental Representations*, 221, Indiana University Press (1988).
13. Kahn, Donald W., *Topology : an introduction to the point-set and algebraic areas*, 11, Dover Publications Inc. (1995).
14. Halmos, Paul R., *Naïve Set Theory*, 56, Springer-Verlag (1974).
15. ibid, p. 66.
16. Hocking, John G. and Young, Gail S., Topology, 6, Dover (1961)
17. ibid, p. 57.
18. ibid
19. ibid, p. 68.
20. Urysohn, P., "Uber metrization der kompakten topologischen Raume," *Math. Ann.* **92**, 275-293 (1924).