Report 83-36
Stanford -- KSL

What's New? A Semantic Definition of Novelty.
Russell Greiner, Michael R. Genesereth,
Jun 1983

Stanford Heuristic Programming Project
Report No. HPP-83-36

June 1983

What's New?
A Semantic Definition of Novelty

Russell Greiner and Michael R. Genesereth

Heuristic Programming Project
Department of Computer Science
Stanford University

This is a slightly expanded version of the paper which appeared in the

Proceedings of IJCAI-83.

1 Introduction

A central process in any learning experience is the incorporation of a *new* fact into an existing theory. Often the goal of that process is more specific, to learn some new fact about some concept. But what does it mean to claim that a sentence is *new*, and even more interesting, what qualifies as a novel fact about some concept? Despite the vast interest in learning and the abundance of related papers (cf. [Dietterich 81a], [Buchanan 78], [Michalski 83], [Dietterich 81b], [Dietterich 82]), no one has rigorously defined what it means to be "new", either in general or with respect to a single concept.

This paper attempts to fill that gap. Our goal is to obtain a *semantic* rather than a *syntactic* understanding of novelty. This preference stems from our belief that a semantic account (one based on the possible interpretations of the theory) provides important insight into the phenomenon of novelty. It also means we may be able to generalize these results to other logics and languages.

The results of this research are relevant (and useful) to many different fields. The primary importance of this work is in providing a first stab at describing the different senses of novelty. In addition to the applications a complete and adequate definition of newness would have as an analytic tool, there are possible applications in knowledge acquisition, representation, and discourse analysis. Many of these stem from the intimate connection between novelty and the intuitive notion of "aboutness". (Section 4 elaborates on each of these.)

This report discusses two kinds of novelty. Section 2 describes newness of a sentence with respect to a theory. Section 3 uses this result to address the more difficult task of determining when a sentence conveys something new about a particular concept. While the first kind of novelty is fairly easy to capture, the second requires a consideration of the interconnections among facts within a theory. Section 4 justifies why this undertaking is relevant and describes how these results may eventually be used. The concluding Section 5 lists several outstanding research issues.

Each of the appendices discusses some relevant digression. In particular, Appendix A motivates and illustrates our intended meaning of this second kind of novelty, by presenting a typical example.

2 New with respect to a Theory

This section addresses the issue of what it means for a sentence σ to be new with respect to a

theory[1] **Th**; this is the relation $N(Th,\sigma)$. Intuitively, we want σ to be new if it (somehow) further specifies something about the world. Alternatively we can think of a new sentence as providing some additional constraints, which remove some possible worlds ([Moore 80]) from consideration.

We first consider a semantic definition of newness: σ is new with respect to a theory if it eliminates some possible interpretation of that theory.[2] That is, given any theory **Th**, in the language **L**, there is a set of models $I^{Th} = \{I_j\}$, in which each I_j maps the symbols of L into objects or sets of tuples of objects in the "real world" in the standard way. That is, each constant is mapped onto an object, and each n-ary relation onto its characteristic set -- those n-tuples that satisfy it. Notice that this means that the universe is fixed beforehand and that these ranges can overlap. (Appendix A elaborates this point.)

Adding additional sentences to a theory can only restrict the set of possible interpretations: $Th \subseteq Th'$ means that $I^{Th'} \subseteq I^{Th}$. This leads to the proposal that

Defn 1: $N_{Sem}(Th,\sigma) \Leftrightarrow I^{Th+\sigma} \subsetneqq I^{Th}$.

This same definition can be expressed syntactically (in terms of logical deducibility rather than semantic validity) using

Defn 2: $N_{Syn}(Th,\sigma) \Leftrightarrow (\sigma \notin Th) \wedge (\neg\sigma \notin Th)$.

As these are equivalent (whenever the language of the theory remains fixed; see [Enderton 72],) we will simply use **N**. (Appendix B sketches a proof for this claim.)

3 New with respect to a Concept (and a Theory)

In many situations it is not enough to realize that an assertion is new; rather, one often wants to claim that it is a new fact *about some concept*. With this in mind, we define the ternary relation $New(Th,s,\sigma)$ to mean that the assertion σ expresses a new fact about the concept **s** with respect to the theory **Th**. Notice we will be dealing with a syntactic symbol, as opposed to its referent. The "learning step" involves adding this sentence σ (along with all of its deductive consequences) to the theory.

What should go into a definition of $New(Th,s,\sigma)$? Clearly, a necessary condition is that σ be a new fact with respect to the entire theory; that is, $N(Th,\sigma)$. But beyond that, we want to capture the

---

[1] We are taking a slightly unorthodox *syntactic* view of theory: viz., a theory is a consistent and deductively closed set of axioms. We will also assume that the deductive system is complete.

[2] We are only concerned with "true interpretations", which map symbols into referents within a *legal* model. We chose "interpretation" rather than "model" to emphasize that we are dealing with a mapping rather than its range.

sense in which σ further specifies the concept **s**, or enables the derivation of additional relevant conclusions about **s**.

This section will present a definition of **New** by proposing a series of "increasingly more nearly correct" descriptions. For simplicity, the examples are taken from propositional logic. (We will later discuss the ways this is, and is not, a limitation.) Each proposal will begin with the intuition involved, followed by a formal description of this conjecture, and finally (in all but the final case) an example of the failure of this conjecture and an analysis of this shortcoming.

Conjecture 1: Syntactic Method. Most statements which relay information about some concept will contain the symbol that refers to that concept. This leads to the proposed syntactic solution: The sentence σ conveys new information about the symbol **A** if the token "A" is lexically included in the string of tokens which form σ, denoted with the assertion **LexInclude("A",σ)** — e.g., **LexInclude("A","A∧B")**. For the reasons mentioned above, we will further insist that **N(Th,σ)**. Formally,

Defn 3: $New_{Syn}(Th,s,\sigma) \Leftrightarrow N(Th,\sigma) \wedge LexInclude(s,\sigma)$.

Unfortunately, this syntactic condition is neither necessary nor sufficient. To see that it is not necessary, realize that we want **New({A⇔B},"A","B")** to be true, since asserting **B** in this situation means that **A** must now be true, which had not been the case before that assertion.

To show insufficiency is a little trickier. Should **New({A∨B},"A","A⇒B")** be true? We argue the answer is no: In this context, asserting **A⇒B** is the same as asserting **B**, which we know says nothing new about **A**. We clearly need a more powerful method for specifying novelty.

Conjecture 2: Fewer Interpretations.

(Subtitle: Model Theory, to the rescue!)

Using the notion of possible interpretations discussed in Section 2, we can define the "interpretation range" of a particular symbol. Let the term $I_j(s)$ designate the "real world" referent of the symbol **s**, given by the interpretation I_j — here, either $\underline{I}$ or $\underline{F}$.[3] We use this to define the interpretation range of the symbol **s**, $I^{Th}(s)$, by

Defn 4: $I^{Th}(s) = \{I_j(s) \mid I_j \in I^{Th}\}$.

As additional facts can only further restrict the range of possible interpretations for any symbol, we have $I^{Th'}(s) \subseteq I^{Th}(s)$ whenever $Th \subseteq Th'$. This leads to our second conjecture,

---

[3] The underbar notation denotes the referent of the corresponding linguistic symbol.

Defn 5: $New_{FI}(Th,s,\sigma) \Leftrightarrow I^{Th+\sigma}(s) \subsetneq I^{Th}(s)$.

This New_{FI} definition seems, at first, adequate. In addition to paralleling the **N** situation, it also resonates nicely with the ideas of Shannon's Information Theory, in which information is tied to the reduction of uncertainty in the distribution of possible values of a signal. (See [Gallager 78].) It also handles the two cases used above to discredit New_{Syn}.

Unfortunately, this New_{FI} requirement does not include all desired cases. There are some sentences that do convey new information in the informal sense outlined above but that do not satisfy this constraint: Start with an empty theory, $ThA \leftarrow \{\}$,[4] in the language $L = \{A, B\}$. The four possible interpretations are shown in Figure 1.

	A	B
I_0	F	F
I_1	F	I
I_2	I	F
I_3	I	I

Figure 1: Interpretations of A and B in ThA.

By inspection, $I^{ThA}(\text{"A"}) = \{I, F\}$. Now, form $ThB \leftarrow ThA + \text{"A}\Leftrightarrow\text{B"}$. While this leaves only two of the four original interpretations, I_0 and I_3, $I^{ThB}(\text{"A"})$ remains $\{I, F\}$, indicating that $A\Leftrightarrow B$ said nothing New_{FI} about **A**.

Although New_{FI} rejects this $A\Leftrightarrow B$, we still believe it should be considered a new fact about **A** in this situation: If we later learn $\neg B$, we will be able to infer that $\neg A$, a conclusion that would not have followed without that sentence. That is, any sentence that makes **A**'s range of interpretations dependent on some other symbol (in the sense that $A\Leftrightarrow B$ made **A** dependent on **B**) also "feels" new.

So there are at least two ways a statement can be new:

- It directly limits the interpretation range of **A**, or
- It establishes (or increases) a dependency of **A** on some other symbols (as that may, in turn, lead to a reduction of the above type).

While New_{FI} covers the first case exactly, it fails in the second.

---

[4] The notation $T \leftarrow \psi$ means the theory **T** is assigned the deductive closure of the set, ψ; and $Th + \sigma$ refers to the deductive closure of $Th \cup \{\sigma\}$.

Conjecture 3: Partial Interpretations. To define dependency requires an understanding of what it means for one symbol to depend on some other symbols. The **A⇔B** case above is clearly one instance of this. In addition to such singular dependencies (of **A** on one other symbol,) **A** may depend on a combination of symbols. (Consider the assertion **A⇔(B⇔C)**. Fixing any assignment to **B** alone, **A** will still be "arbitrary"; that is, it could be either T or F, depending on **C**. However, if both **B** and **C** are fixed, then **A** is fully determined.)

We saw that σ is a new fact about **A** if it increases **A**'s dependency on some n-tuple of symbols $\langle s_1, \ldots, s_N \rangle$, that is, if asserting σ restricts the set of assignments available to **A**, given some assignment $\langle \underline{S}_1, \ldots, \underline{S}_N \rangle$ to the symbols $\langle s_1, \ldots, s_N \rangle$. For example, we noted that **A** was more dependent on **B** in **ThB** than in **ThA**. We see this by considering the assignment of **B** to F. In **ThA**, **A**'s value could be either F or T, independent of this assignment to **B**. However, **A** can no longer be assigned T in **ThB**, given this assignment to **B**; its value is now restricted to F.

As this "$\langle s_1, \ldots, s_N \rangle$ assignment to $\langle \underline{S}_1, \ldots, \underline{S}_N \rangle$" reflects an assignment of only a subset of the symbols, $\{s_1\} \subseteq \mathbf{L}$, we will call it a *partial interpretation*. We can associate with each partial interpretation the equivalent class of full interpretations that agree on the assignment of a set of symbols. Formally, take any function that maps some of the symbols of the language into the universe, **U** — that is, any $\varphi : \xi \mapsto \mathbf{U}$, where $\xi \subseteq \mathbf{L}$. We can use this to define the equivalence class $[\![\varphi^{Th}]\!]$ as the set of interpretations that consistently extend φ — that is, it includes each interpretation that agrees with φ's assignment of each symbol in φ's domain and assigns every other element of **L** in some consistent manner.

Defn 6: $[\![\varphi^{Th}]\!] = \{I \in \mathbf{I}^{Th} \mid \forall x \in \text{Domain}[\varphi].\ \varphi(x) = I(x)\}$.

The assignments of **s** that are consistent with the partial interpretation $[\![\varphi^{Th}]\!]$ are just

Defn 7: $[\![\varphi^{Th}]\!](s) = \{I(s) \mid I \in [\![\varphi^{Th}]\!]\}$.

With this notation, we can state that **A**'s dependency on the symbols $\xi = \mathbf{Domain}[\varphi]$ increases if the set of possible values of **A** consistent with the partial interpretation, $[\![\varphi^{Th}]\!]$**("A")**, decreases but remains non-empty. (Seeing it vanish means that there are no values of **A** that are consistent with this assignment, which means that no models can be derived by extending this partial interpretation.)

To test if σ is new, therefore, consider all of the assignments available to **A** for each partial interpretation, before and after adding this purportedly new sentence σ to the theory. If the number of possible referents of **A** decreases for any partial interpretation (and remains non-empty), we will declare that σ is new.

Defn 8: $\mathbf{New_{PI}}(Th, s, \sigma) \Leftrightarrow \exists \varphi.\ [\![\varphi^{Th+\sigma}]\!](s) \subsetneq [\![\varphi^{Th}]\!](s) \wedge [\![\varphi^{Th+\sigma}]\!](s) \neq \{\}$

A few notes:

1. We will say that the particular function φ whose partial interpretation $[\![\varphi^{Th}]\!]$ decreased in the above equation is a "witness" to σ's novelty (with respect to **A**).

2. This definition subsumes the $\mathbf{New_{FI}}(Th, A, \sigma)$ condition. This follows from the fact that $[\![\{\}^{Th}]\!](A)$ is equal to $I^{Th}[A]$. (Note this {} mapping is the "null mapping", whose domain is empty.)

3. Realize that if $A \in \mathbf{Domain}[\varphi]$, then $[\![\varphi^{Th}]\!](A)$ would contain a single member. As there are no non-empty proper subsets of such singleton sets, it is sufficient to use $\mathbf{Domain}[\varphi] \subseteq L - \{A\}$, rather than $\mathbf{Domain}[\varphi] \subseteq L$.

4. Consider the set of "almost complete interpretations" $[\![\varphi^{Th}]\!]$, whose domain includes every symbol of **L** except **A**; that is, $\mathbf{Domain}[\varphi] = L - \{A\}$. While it may appear that these partial interpretations are sufficient — that one of these would witness any new σ — the counterexample below shows that is not the case.

A tableau helps to visualize this definition. The left tableau in Figure 2 corresponds to the theory **ThC** ← {**A**⇔**B**} in the language **L** = {**A**,**B**,**C**}, and the one on the right to **ThD** ← **ThC** + "**A**⇔**C**". Each row is indexed by a mapping φ and each column by an assignment to **A**. A tableau position is tagged with a "1" if this assignment of $\mathbf{Domain}[\varphi] \cup \{A\}$ is consistent — that is, if there is any full interpretation associated with this position — and a "0" otherwise.

Finding a witness to a sentence's novelty reduces to finding a row, r, in which a "1" is flipped to "0" but which does not vanish — that is, r must retain a "1" in some position. The fifth and sixth rows below (labeled with the "{<C,F>}" and "{<C,T>}" mappings) each satisfy this property, showing that $\mathbf{New_{PI}}$(**ThC**,"**A**","**A**⇔**C**"). Notice that none of the top four rows, which correspond to those "almost complete interpretations" has that property, demonstrating the point of Item 4 above. (These rows do form an adequate spanning set, though, as all the other rows can be derived by ORing together appropriate sets of these.)

4 Uses

Even in its current unmechanized form, this definition can be used effectively as an analytic tool with which to understand many existing learning programs. Eventually, we hope to develop a "NewP" predicate or possibly a pair of operational (multivalue) functions: "Newσ", which generates $\mathbf{New_{PI}}$ sentences from a given theory and symbol, and "NewSYM", which returns the symbols for which a given sentence is $\mathbf{New_{PI}}$. This section lists several ways this definition (and its operationalizations) can be used.

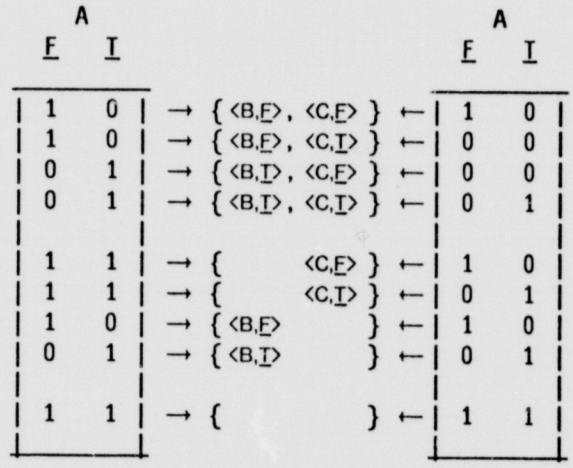

Figure 2: Partial interpretations for ThC and ThD.

- *Analytic Tool.* An adequate definition of newness would help us identify the sources (and recipients) of novelty within learning programs. For example, the teacher provides the ARCH program ([Winston 75]) with the new facts that enable it to learn. LEX's problem solver and critic are the sources of novelty for the rest of the system ([Mitchell 81]). AM ([Lenat 82]) has no clear source of novelty. This definition may also help us understand the distinction between compositional new terms — such as AM's definition of prime numbers — and other new terms, such as Bacon's use of intrinsic properties ([Langley 79]). Finally, it may lead to a definition of learning not based exclusively on performance.
- *Learning and Knowledge Acquisition.* An adequate (i.e., computable) definition of novelty might suggest ways of learning a topic more effectively. For example, it could focus the learner's efforts on those aspects of the domain where he has the greatest potential for acquiring something new. This information would help a knowledge-base builder decide which concepts need to be better understood, helping him to direct the dialogue. An analysis of a symbol's dependencies (defined above) might then be used to generate appropriate "probe" sentences to help understand this still vague concept.
- *Representation.* How should a given proposition be indexed? In general each concept should point to all the relevant facts that are *about* that concept. The most obvious approach, based strictly on lexical inclusion, is inadequate. For example, one would want to index "x + 1 = 0" by "x" and not by " + ", whereas "x + y = y + x" should be associated with " + " and not with "x".

  So how does one determine those concepts that a given fact is really about? We claim that "aboutness" is intimately tied to "newness" in the sense that σ is *about* a concept **c** whenever this σ expresses something *new* about **c** with respect to the appropriate diminished theory (which excludes g[s] and all of its consequences).
- *Linguistics.* The basic purpose of communication is for the speaker, S, to transmit a set of *new* facts, usually about some specific topic. To understand this process, we have to know what it means for a fact to be new to H and then how S (and H) can use this meta-fact when constructing (or understanding) the message.[5]

[5] This research stemmed from the authors' efforts to understand a particular type of communication, analogy, in terms of this model.

5 Conclusion

While we have yet to solve all the issues associated with this model of novelty, this paper would be incomplete if it did not address the following topics:

- *Applicability.* The $\mathbf{New}_{PI}$ relation described above is applicable to any symbol in predicate calculus as well as propositional logic. In particular, the same formalism we saw work for constant symbols works adequately for relation symbols, albeit with an even larger tableau.
- *"Assertional Novelty".* The novelty we discussed above, $\mathbf{New}_{PI}$, is "definitional", in that its goal is to specify more precisely the referent of a given symbol. Another source of novelty comes from specifying some attribute of the concept; we label such facts "assertionally novel". (See [Woods 75].)

  These two categories are distinct: Imagine the symbol **RDG** had been totally determined, in the sense that the set $\mathbf{I}^{\mathbf{Th}}$**("RDG")** had but a single member. As such, there is nothing $\mathbf{New}_{PI}$ we can say about **RDG**. Despite this certainty, you still might not know what his hair color is. That is, **HairColor(RDG Brunette)** might be true in one interpretation, whereas others might hold that **HairColor(RDG Blond)**. Clearly **HairColor(RDG Blond)** is a $\mathbf{New}_{PI}$ fact about **HairColor**; however, most people would also want this it to be a new fact about **RDG** as well — that is, $\mathbf{New}_{\mathbf{Assert}}$**(Th,"RDG","HairColor(RDG Blond)")**.
- *Intensional, not Extensional.* This paper has dealt exclusively with extensional phenomena, where novelty was determined with respect to the extensions of the symbols. Another approach is intensional — based on descriptions.
- *Deductively Closed.* Probably the most serious criticism of this work is its dependency on a complete deductive system and the requirement that each theory be deductively closed. New-sounding statements can also be used to focus the hearer's attention on some facts he already knew, rather than expose him to new facts. It should be possible to extend this formalism to handle such resource-limited deducibility. Then we could address topics like monotonic novelty and information obsolence.

Each of the issues mentioned above suggests a research task — that of plugging each limitation. The two issues we find most pressing are:

- Finding an equivalent but syntactical formulation of the semantical $\mathbf{New}_{PI}$ relation, in the same manner that $\mathbf{N}_{Syn}$ matched $\mathbf{N}_{Sem}$. We hope this will lead to one or more operational versions, of the types mentioned in the beginning of Section 4.
- Expanding this $\mathbf{New}_{PI}$ definition to work with deductive systems that are incomplete. (This reiterates the last issue shown above.)

Our basic thesis is that σ is a new fact about **A**, with respect to the theory **Th**, if, *under some set of circumstances, σ limits the number of interpretations of* **A**. $\mathbf{New}_{PI}$ achieves this by examining every partial interpretation, testing each to see if **A** loses a possible interpretation in that situation. This "partial interpretation" definition of context is clearly as general as possible. Furthermore, by counterexample, we have shown that this extreme generality is necessary.

These issues all seem fertile ground for further investigation. Such explorations may very well lead to interesting and usable new results.

Acknowledgments

We would especially like to thank Tom Dietterich for his significant contributions to this work. The following people also contributed: Professor Bruce Buchanan, Dr. Lew Creary, Jim Davidson, Dr. Johan deKleer, Dianne Kanerva Dr. Jussi Ketonen, Jock Mackinlay, Dr. Robert Moore, Yoram Moses, and Ben Moszkowski. Many useful comments were provided by the reviewers. This basic research was funded by ARPA Contract #MDA903-80-C-0107.

Bibliography

[Buchanan 78] Buchanan, B. G., Mitchell, T. M., Smith, R. G. and Johnson, C. R. Jr.
Models of Learning Systems.
In *Encyclopedia of Computer Science and Technology*, . Dekker, 1978.

[Dietterich 81a] Dietterich, T. G. and Buchanan, B. G.
The Role of the Critic in Learning Systems.
Technical Report STAN-CS-81-891, Computer Science Department, Stanford University, December, 1981.

[Dietterich 81b] Dietterich, T. G. and Michalski, R. S.
Inductive Learning of Structural Descriptions.
Artificial Intelligence 16, 1981.

[Dietterich 82] Dietterich, T. G., London, R., Clarkson, K., and Dromey, G.
Learning and Inductive Inference.
In Cohen, P. and Feigenbaum, E.A. (editors), *The Handbook of Aritificial Intelligence*, . William Kaufman, Inc., Los Altos, CA, 1982.

[Enderton 72] Enderton, Herbert B.
A Mathematical Introduction to Logic.
Academic Press, Inc., New York, 1972.

[Gallager 78] Gallager, Robert G.
Information Theory and Reliable Communication.
John Wiley and Sons, Inc., New York, 1978.

[Langley 79] Langley, Pat.
Rediscovering Physics With BACON.3.
In *6-IJCAI*, pages 505-507. Tokyo, August, 1979.

[Lenat 82] Lenat, Douglas B.
AM: Discovery in Mathematics as Heuristic Search.
In Davis, Randall and Lenat, Douglas B. (editors), *Knowledge-Based Systems in Artificial Intelligence*, . McGraw-Hill International Book Company, San Francisco, 1982.

[Michalski 83] Michalski, Ryszard S., Carbonell, Jaime G., and Mitchell, Tom M. (editors).
Machine Learning: An Artificial Intelligence Approach.
Tioga Publishing Company, Palo Alto, CA, 1983.

[Mitchell 81] Mitchell, Thomas M., Utgoff, Paul E., Nudel, Bernard and Banerji, Ranan.
Learning Problem-Solving Heuristics thruough Practice.
In *IJCAI-7*, pages 127-134. UBC, 1981.

[Moore 80] Moore, Robert C.
Reasoning about Knowledge and Action.
Technical Note 191, SRI Interational, October, 1980.

[Winston 75] Winston, P. H.
Learning Structural Descriptions from Examples.
McGraw-Hill Book Company, New York, 1975, chapter 5.

[Woods 75] Woods, W. A.
What's in a Link: Foundations for Semantic Networks.
In D. G. Bobrow & A. M. Collins (editors), *Representation and Understanding*. Academic Press, 1975.

A. A Typical Novelty Situation

This appendix presents a series of "novelty situations", designed to motivate and illustrate our intended meaning of the ternary form of novelty, **New(Th,c,σ)**. In each of these, we argue why a certain fact feels "new" with respect to a particular concept, in the context of a particular theory. These examples, expressed in predicate calculus, parallel the propositional logic ones given in the text. They also indicate the assumptions which underlie our intended use of semantic interpretations.

Begin with an initial theory which contains the following three facts:

$Th_0 \leftarrow$ { (Person RDG)
(Article WN)
(∀ $P,$A. [(Author $A $P) ∧ (Laconic $A)]
⇒ (ShortPaper $P) }

(It may include other sentences as well; but nothing which relates to either RDG or WN.)

Now augment this theory by adding the sentence (In WN IJCAI-83), to form

$Th_1 \leftarrow Th_0$ + "(In WN IJCAI-83)".

This fact seems new with respect to WN, but not with respect to RDG. We can express this by

¬New(Th_0, "RDG", "(In WN IJCAI-83)")
New(Th_0, "WN", "(In WN IJCAI-83)")

Later we learn that (Author WN RDG), building

$Th_2 \leftarrow Th_1$ + "(Author WN RDG)".

This "authorship" statement seems like a new fact about both WN and RDG — hence,

New(Th_1, "WN", "(Author WN RDG)")
New(Th_1, "RDG", "(Author WN RDG)").

We are then told that

$Th_3 \leftarrow Th_2$ + "(Laconic RDG)".

Not only does this express something about RDG, but, using the "laconic people write short papers" rule above, we can deduce that (ShortPaper WN). For this reason we want both

New(Th_2, "RDG", "(Laconic RDG)")
New(Th_2, "NW", "(Laconic RDG)")

Finally we form

$Th_4 \leftarrow Th_3$ + "(Topic WN (Topic WN1))".

We will discuss how this sentence is novel, later.

Some comments are in order.

- A given sentence may be new with respect to one concept and not new with respect to others

(in the context of the same theory):

$\neg$New(Th_0, "RDG", "(In WN IJCAI-83)") and
New(Th_0, "WN", "(In WN IJCAI-83)").

- The same sentence can be new with respect to several distinct concepts:

New(Th_1, "WN", "(Author WN RDG)") and
New(Th_1, "RDG", "(Author WN RDG)").

- A given sentence may be new with respect to a concept in the context of one theory but not new, with respect to the same concept, in the context of another theory:

New(Th_0, "WN", "(In WN IJCAI-83)") but
$\neg$New(Th_1, "WN", "(In WN IJCAI-83)").

- A sentence may be new with respect to a symbol not included in that sentence:

New(Th_2, "WN", "(Laconic RDG)").

(The conclusion of Section 3's first conjecture reiterates this point.)

The defination of New_{FI}, (which is discussed in the second conjecture in Section 3,) is based on the number of interpretations of the symbol in question. For this to make any sense we have to assume (1) that the models involved "overlap", in the sense that each interpretation maps symbols into the same range of possible elements, and (2) that some symbols of the language are "fixed" — that is, their respective extensions are understood and unambiguous. Realize we have implicitly assumed this all along: in this example we "knew" the desired referent of the constant symbol **IJCAI-83** and the predicate symbol **Person**. **IJCAI-83** obviously refers to a particular book; or, stated another way, it designates the same "thing" in every model. Similarly **(Person RDG)** means that **RDG**'s denotation is a person.[6] Given that these symbols, among others, are fixed, we can build the table shown in Figure A.

That is, all Th_0 knew about **RDG** was that **(Person RDG)**, which means that **RDG** could refer to any of the 4.5 billion people in the world. Similarly **WN** could be any of the world's 10 million articles. (As mentioned above, this is only meaningful if the concept **Person** (resp. **Article**) is well defined.)

On hearing that **(In WN IJCAI-83)**, we know that **WN** must be one of only about 250 articles in the IJCAI-83 proceedings. This "10 million to 250" reduction implies that **(In WN IJCAI-83)** is a New_{FI} fact about **WN**. As $\|I^{Th1}(\text{"RDG"})\| = \|I^{Th2}(\text{"WN"})\| = 4.5E9$, we see that

---

[6] This "some symbols are fixed" assumption is not as major restriction as it may seem. In fact, this is the typical of most communication situations. Usually both speaker and hearer have a common "consensual domain" of understanding, in which most concepts are unambiguous and are commonly understood — as the predicate **Person** is in day-to-day discourse. The typical goal of these communication acts is to relay a (new) fact about some other, less-well-defined concept, defining it in terms of those well-understood concepts.

th	*facts*	$\|\|I^{th}(\text{"RDG"})\|\|$	$\|\|I^{th}(\text{"WN"})\|\|$
	(Article WN) (Person RDG) ∀ $P,$A. IF [(Author $P $A) ∧ (Laconic $A)] THEN (ShortPaper $P) ...		
Th_0		4.5E9	1E7
Th_1	(In WN IJCAI-83)	4.5E9	300
Th_2	(Author WN RDG)	300	250
Th_3	(Laconic RDG)	15	10
Th_4	(Topic WN (Topic WN1))	15	10

Figure 3: Number of Interpretations of RDG and WN, in various Theories

(In WN IJCAI-83) was not a $\mathbf{New_{FI}}$ fact about RDG, as desired.

We similarly see that

New(Th_1, "RDG", "(Author WN RDG)").
New(Th_2, "RDG", "(Laconic RDG)")
New(Th_2, "WN", "(Laconic RDG)")

Unfortunately this analysis leads to the conclusion that $\neg$**New$_{FI}$(Th$_3$,"WN","(Topic WN (Topic WN1))")**, which does NOT correspond to our intuitions: While this statement does not restrict the number of interpretations of the symbol, (here **WN**,) it does establish a dependency of this concept on another, (here **WN1**). We feel that statements like this, which establish a dependency of one concept on another, should be considered novel. Our argument is that later, when we learn that **(Topic WN1 Novelty)**, we will be able to infer **(Topic WN Novelty)**, which in turn means that **WN** is then totally specified — that is, that it must refer to one specific article, the original version of the one you are reading now.

Hence, we see that $\mathbf{New_{FI}}$ is slightly inadequate, as there are some sentences which our intuitions claim should be deemed novel, which it declares are not. This leads to the $\mathbf{New_{PI}}$ relation, defined in

Conjecture 3 of Section 3.

B. Equivalence of N_{Syn} and N_{Sem}

This appendix sketches the proof that N_{Syn} and N_{Sem} are equivalent.

$N_{Syn}(Th,\sigma) \Rightarrow N_{Sem}(Th,\sigma)$:

Define $Th' \leftarrow Th+\sigma$. As $I_{Th'} \subseteq I_{Th}$, it is sufficient to show that $I_{Th'} \neq I_{Th}$. Assume that σ did not eliminate any member of I_{Th}; i.e. that this σ "just happened" to be true under all of those interpretations. That would mean that σ is true in all models of Th; i.e. $Th \models \sigma$. This means that $Th \vdash \sigma$, by Godel's Completeness Theorem. But this directly contradicts the initial assumption that $Th \nvdash \sigma$, from $N_{Sym}(Th,\sigma)$. Hence there must be some interpretation, I_j, in which this σ is false; and this I_j must therefore be thrown out of $I_{Th'}$, proving that $I_{Th'} \neq I_{Th}$, as desired.

$N_{Sem}(Th,\sigma) \Rightarrow N_{Syn}(Th,\sigma)$:

As $I_{Th'} \neq I_{Th}$, clearly $Th' \neq Th$, which means that $\sigma \notin Th$. Had $\neg\sigma \in Th$, then Th' would be inconsistent. $I_{Th'}$ would either be undefined, or include every possible interpretation (if you insist); and, in either case, not be contained within I_{Th}. ∎

C. Preliminary Connection to Syntactic Methods

Before we can begin to mechanize this novelty relation, we must first express it in a syntactic form. This chapter will present a first shot at this proof-based formulation, for the proposition case. (We feel that some parts of this will scale up, *mutatis mutandis*, to the full predicate calculus case.)

Recall that $New_{PI}(Th,s,\sigma)$ is true when some interpretation of the symbol s, which was possible in some context of the theory Th, is ruled out in the enhanced theory $Th+\sigma$, within the corresponding, still-possible context.

That is,

$$New_{PI}(Th,s,\sigma) \Leftrightarrow$$
$$\exists\, \rho,\alpha.\ (Th \nvdash \rho \Rightarrow \neg\alpha(s)) \wedge$$
$$(Th,\sigma \vdash \rho \Rightarrow \neg\alpha(s)) \wedge$$
$$(Th,\sigma \nvdash \neg\rho)$$

where ρ is a sentence and α is a unary prediate. WLOG this ρ can be a conjunct of assignments to various proposition symbols, i.e. each term is either a symbol or its negation. As such we can think of it as referring to some row of the tableau. (If only expressibility was so trival in the predicate calculus case...)

Returning to the $\mathbf{New}_{PI}(\mathbf{ThC}, \text{"A"}, \text{"A}\Leftrightarrow\mathbf{C}\text{"})$ example used in Conjecture 3 of Section 3, we see that setting

$\mathbf{Th}$	$\equiv$	$\{ \mathbf{A}\Leftrightarrow\mathbf{B} \}$	
ρ	$\equiv$	$\neg\mathbf{C}$	— describing the fifth row of the tableaus of Figure 2
$\alpha(\mathbf{s})$	$\equiv$	$\mathbf{s}$	— meaning that $\alpha(A) = A$

leads to the conclusion that $\sigma = \mathbf{A}\Leftrightarrow\mathbf{C}$ is indeed a $\mathbf{New}_{PI}$ fact about **A** in this theory.

Proof of equivalency to the semantic form:

The first two clauses, $\mathbf{Th} \nvdash (\rho \Rightarrow \neg\alpha(\mathbf{s}))$ and $\mathbf{Th}, \sigma \vdash (\rho \Rightarrow \neg\alpha(\mathbf{s}))$, mean that there is some interpretation of **s**, expressed by $\alpha(\mathbf{s})$, which is possible given the theory **Th**, within the context defined by ρ. In this propositional case we can syntactically specify any partial interpretation (read "row of the tableau",) as a simple sentence.

As there is something NOT derivable from $\mathbf{Th} + \sigma$, that larger theory must be consistent; meaning that **Th** must have been consistent, and σ not contradictory. The final clause means that ρ is still satisfiable — i.e. that that row has not vanished. ∎

Unfortunately, the predicate calculus case is not nearly so simple. First, there are other types of novelty we must worry about here. (Recall that the only type of (extensional) novelty in this propositional case is this definitional case.) Furthermore, even for this definitional situation is complicated by the issue of expressibility: unlike propositional logic, where any interpretation can be syntactically expressed, there are "partial interpretations" which cannot be expressed in predicate calculus.