Between Chaos and Structure: Interpreting Lexical Data through a Theoretical Lens

James Pustejovsky
Dept. of Computer Science
Brandeis University
Waltham, MA USA
jamesp@cs.brandeis.edu

Anna Rumshisky
Dept. of Computer Science
Brandeis University
Waltham, MA USA
arum@cs.brandeis.edu

Abstract

In this paper, we explore the inherent tension between corpus data and linguistic theory that aims to model it, with particular reference to the dynamic and variable nature of the lexicon. We explore the process through which modeling of the data is accomplished, presenting itself as a sequence of conflicting stages of discovery. First-stage data analysis informs the model, whereas the seeming chaos of organic data inevitably violates our theoretical assumptions. But in the end, it is restrictions apparent in the data that call for postulating structure within a revised theoretical model. We show the complete cycle using two case studies and discuss the implications.

1 Introduction

This paper is an attempt to demonstrate both the theoretical significance of data and the empirical significance of theory. In short, it is an essay on the relationship between data and theory in linguistics. We begin by examining the role theory plays in the analysis of language.

Within analytic linguistics, our initial assumptions on what structures should exist in the data provide us with predictive force and guide us through an often muddled and contradictory set of facts requiring analysis. This initial stage of investigation, what we could call first-level data analysis, uses phenomenological data that are constructed by matching words to expressions predicted by analytic grammar. We refer to these as *synthetic data*. When real data do not fit, we add new structure or, just as often, we idealize such data away. In fact, it could be argued that no data analysis is ever performed without some theoretical bias. This approach has been the operative standard in most language analysis since Aristotle. Except of course, within corpus linguistics, to which we turn next.

Contrary to analytic linguistics, corpus linguists and lexicographers have long stressed the role that extensive analysis of text in use plays in any language description [Sinclair, 1991, Firth, 1957, Hornby, 1954]. Such work emphasizes the importance of looking at the data without theoretical pruning (cf. [Sinclair, 1966, Sinclair, 2004, Hanks, 1994, Hanks, 1996], among others) The basis of this approach is the examination of *organic data* (as opposed to synthetic) in order to form hypotheses regarding language and linguistic behavior. Lexicographic studies of concordancing and collocations have long been used as a means for examining the data [Sinclair, 1987, Church and Hanks, 1990].

In this paper, we use a corpus-driven approach to test a theoretically motivated first-level data analysis of two linguistic phenomena. We will see that theory predicts behavior that is not attested, and behavior exists that is not predicted by theory. These data will be used to inform and update the theory, and in some cases modify or drop theoretical assumptions. The resulting process is an interplay of data analysis and theoretical description.

2 Theoretical Preliminaries

For this exercise in how theory and data interact, we adopt the model of Generative Lexicon, a theory of linguistic semantics which focuses on the distributed nature of compositionality in natural language [Pustejovsky, 1995]. Unlike purely verb-based approaches to compositionality, Generative Lexicon (henceforth, GL) attempts to spread the semantic load across all constituents of the utterance. Overall, GL is concerned with explaining the creative use of language; we consider the lexicon to be the key repository holding much of the information underlying this phenomenon. More specifically, however, it is the notion of a constantly evolving lexicon that GL attempts to emulate; this is in contrast to currently prevalent views of static lexicon design, where the set of contexts licensing the use of words is determined in advance, and there are no formal mechanisms offered for expanding this set.

Traditionally, the organization of lexicons in both theoretical linguistics and natural language processing systems assumes that word meaning can be exhaustively defined by an enumerable set of senses per word. Lexicons, to date, generally tend to follow this organization. As a result, whenever natural language interpretation tasks face the problem of lexical ambiguity, a particular approach to disambiguation is warranted. The system attempts to select the most appropriate 'definition' available under the lexical entry for any given word; the selection process is driven by matching sense characterizations against contextual factors. One disadvantage of such a design follows from the need to specify, ahead of time, all the contexts in which a word might appear; failure to do so results in incomplete coverage. Furthermore, dictionaries and lexicons currently are of a distinctly static nature: the division into separate word senses not only precludes permeability; it also fails to account for the creative use of words in novel contexts.

GL attempts to overcome these problems, both in terms of the expressiveness of notation and the kinds of interpretive operations the theory is capable of supporting. Rather than taking a 'snapshot' of language at any moment of time and freezing it into lists of word sense specifications, the model of the lexicon proposed here does not preclude extensibility: it is open-ended in nature and accounts for the novel, creative, uses of words in a variety of contexts by positing procedures for generating semantic expressions for words on the basis of particular contexts. To accomplish this, however, entails making some changes in the formal rules of representation and composition. Perhaps the most controversial aspect of GL has been the manner in which lexically encoded knowledge is exploited in the construction of interpretations for linguistic utterances. Both lexical items and phrases encode the following four types of information structures:

(1) a. LEXICAL TYPING STRUCTURE: giving an explicit type for a word positioned within a type system for the language;
b. ARGUMENT STRUCTURE: specifying the number and nature of the arguments to a predicate;
c. EVENT STRUCTURE: defining the event type of the expression and any internal event structure it may have, with subevents;
d. QUALIA STRUCTURE: a structural differentiation of the predicative force for a lexical item.

The qualia structure, inspired by [Moravcsik, 1975] interpretation of the *aitia* of Aristotle, are defined as the modes of explanation associated with a word or phrase in the language, and are defined as follows [Pustejovsky, 1991]:

(2) a. FORMAL: the basic category which distinguishes the meaning of a

word within a larger domain;
b. CONSTITUTIVE: the relation between an object and its constituent parts;
c. TELIC: the purpose or function of the object, if there is one;
d. AGENTIVE: the factors involved in the object's origins or "coming into being".

The different aspects of lexical meaning listed in (1) and (2) can be packaged together as a set of features, illustrated below, where ARGSTR refers to the argument structure of a predicate and EVENTSTR to the event structure (cf. [Bouillon, 1997, Pustejovsky, 1995])

$$\begin{bmatrix} \alpha & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = x \\ \ldots \end{bmatrix} \\ \text{EVENTSTR} = & \begin{bmatrix} \text{E1} = e_1 \\ \ldots \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{CONST} = \textbf{what } x \textbf{ is made of} \\ \text{FORMAL} = \textbf{what } x \textbf{ is} \\ \text{TELIC} = \textbf{function of } x \\ \text{AGENTIVE} = \textbf{how } x \textbf{ came into being} \end{bmatrix} \end{bmatrix}$$

When certain features (qualia) are present or absent, we can abstract away from the representation, and generalize lexemes as belonging to one of three conceptual categories [Pustejovsky, 2001, Pustejovsky, 2006].

(3) a. NATURAL TYPES: Natural kind concepts consisting of reference only to Formal and Constitutive qualia roles; e.g., *tiger*, *river*, *rock*.
b. ARTIFACTUAL TYPES: Concepts making reference to Telic (purpose or function), or Agentive (origin); e.g., *knife*, *policeman*, *wine*.
c. COMPLEX TYPES: Concepts integrating reference to the relation between types from the other levels; e.g., *book*, *lunch*, *exam*.[1]

This enriched inventory of types for the language is motivated by the need for semantic expressiveness in lexical description. We also need, however, richer interpretive operations to take advantage of these new structures. Following [Pustejovsky, 2006], we argue that there are four ways a predicate can combine with its argument:

[1] That is, *book* can refer both to the information contained in the book and to the physical object, *lunch* can refer both to the event and to the food, etc. For an inventory of complex types, see [Rumshisky et al., 2007]

(4) a. PURE SELECTION (Type Matching): the type a function requires is directly satisfied by the argument;
b. ACCOMMODATION: the type a function requires is inherited by the argument;
c. TYPE COERCION: the type a function requires is imposed on the argument type. This is accomplished by either:

i. *Exploitation*: taking a part of the argument's type to satisfy the function;
ii. *Introduction*: wrapping the argument with the type required by the function.

These mechanisms will form the theoretical scaffolding with which we will perform our first-level data analysis of the argument selection phenomena in the next section. Natural types (e.g. *lion*, *rock*, *water*) are viewed essentially as atomic from the perspective of selection. Conversely, artifactual (or tensor) types (e.g. *knife*, *beer*, *teacher*) have an asymmetric internal structure consisting of a *head type* that defines the nature of the entity and a *tail* that defines the various generic explanatory causes of the entity of the head type. Head and tail are unified by a type constructor $\otimes$ ('tensor') which introduces a qualia relation to the head type: for example, *beer* = $liquid \otimes_{Telic} drink$. That is, *beer* is a kind of liquid; not all liquids are for drinking, but the very purpose (Telic) of *beer* is that someone should drink it.

Finally, complex types (or dot objects) (e.g. *school*, *book*, *lunch* etc.) are obtained through a complex type-construction operation on natural and artifactual types, which reifies two elements into a new type. Dot objects are to be interpreted as objects with a complex type, not as complex objects. The constituents of a complex type pick up specific, distinct, even incompatible aspects of the object. For instance, *lunch* ($event \bullet food$) picks up both $event$ and $food$ interpretations, *speech* ($event \bullet info$) picks up both $event$ and $info$ interpretations, etc. [Asher and Pustejovsky, 2006].

Type exploitation occurs when a verb selects only a part of the semantics associated with its arguments. For example, the verb *buy* selects for a physical object, which is only a part of the dot object $phys \bullet info$ in (5) below:

(5) Mary bought a book.

Type introduction is the converse, where a new structure is wrapped around a type in argument position. Consider the verb *read*, which selects for the aforementioned type $phys \bullet info$ in direct object position. When, for example, an informational noun such as *rumour* appears, it is "wrapped" with the additional type information:

(6) Mary read a rumour about you.

That is, this rumour is not just an idea (proposition) but has physical manifestation, by virtue of type introduction coercion.

3 Data informs Analysis: Two Case Studies

3.1 First-level Data Analysis: Formulating Theoretical Predictions by Introspection

When initially modeling a particular linguistic phenomenon or pattern, the typical linguistic assumption is to "idealize" the data using introspective or phenomenological data. This has become de rigueur in theoretical linguistic investigations, and we will refer to this stage as *first-level data analysis*. Corpus-oriented linguists have long criticized this approach as armchair lexicography (cf. [Fillmore, 1991], Sinclair, Hanks, and others). While they do produce a partial account of the data, reflecting valid observational tendencies, such approaches tend to give an in-depth account of a limited set of behaviors, and typically leave unaccounted the full range of combinatorial phenomena.

Below, we present two case studies in composition: argument selection; and type coercion. We begin by giving a theoretical account of these phenomena, using synthetic data. We then examine the same phenomena using organic data taken from corpora. Finally, we show how the theoretical model of the data is enriched by accounting for a fuller range of the phenomena.

3.1.1 Case Study 1: Verbs Selecting for Artifactual Entities

We begin our investigation with the behavior of verbs that select for artifactual arguments, as defined in the previous section. The theory makes a distinction between natural kinds and non-natural kinds, and this is realized in the types used by the lexicon and the grammar. As a result, verbs will be also be typed as natural and non-natural predicates, depending on what kind of arguments they select for. Hence, *Natural predicates* will be those properties and relations selecting for natural types, while *Artifactual* *predicates* will select for an Artifactual. This distinguishes the classes of verbs in (7) below.

(7) a. NATURAL PREDICATES: touch, sleep, smile
b. ARTIFACTUAL PREDICATES: fix, repair, break, mend, spoil

These classes are defined by the type assigned to the arguments. For example, the type structure for the Natural predicate *touch* is shown in (8):

(8)
$$\begin{bmatrix} \textbf{touch} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = x : \mathit{phys} \\ \text{ARG2} = y : \mathit{phys} \end{bmatrix} \end{bmatrix}$$

An Artifactual predicate such as the verb *repair* would be typed as shown in (9).

(9)
$$\begin{bmatrix} \textbf{repair} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = x : human \\ \text{ARG2} = y : \mathit{phys} \otimes_{Telic} \alpha \end{bmatrix} \end{bmatrix}$$

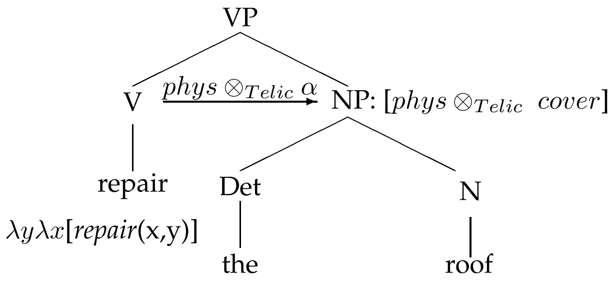
Given these theoretical assumptions, what we expect to encounter as the direct object of artifactual predicates such as *repair*, *fix*, and so forth, are entities that are themselves artifacts.

(10) a. Mary repaired the roof.
b. John fixed the computer.
c. The plumber fixed the sink.
d. The man mended the fence.

What this also predicts is the absence of verb-argument pairings with entities that are not artifactual in some sense. This would appear to be borne out as well, upon initial reflections. You do repair manufactured objects like roofs, cars, and windows; you don't repair natural kinds like boulders, rivers, trees, and pumas.

To illustrate just how this selection is accomplished, consider the sentence in (10a). The verb *repair* (under the intended sense) is typed to select only Artifactual entities as its internal argument. The NP *the roof* satisfies this constraint, as it has a Telic value (i.e., it's an Artifactual), and the verb-argument composition proceeds without incident.

(11)


VP
V
$phys \otimes_{Telic} \alpha$
NP: $[phys \otimes_{Telic}\ cover]$
repair
$\lambda y \lambda x[$*repair*(x,y)$]$
Det
the
N
roof

What this illustrates is how verbs are strictly typed to select specific classes of arguments, in this case an Artifactual as direct object. We consider a somewhat different compositional context in the next section.

3.1.2 Case Study 2: Verbs Selecting for Propositions

As our second study, we examine another aspect of GL's theory of selection, namely the phenomenon of *type coercion*. As we saw in the previous section, *Matching* or *Pure Selection* takes place when the type requested by the verb is directly satisfied by the argument. In this case, no type adjustment is needed. *Accommodation* occurs when the selecting type is inherited through the type of the argument. *Coercion* takes place when there is a mismatch (type clash) between the type requested by the verb and the actual type of the argument. This type clash may trigger two kinds of coercion operations, through which the type required by the function is imposed on the argument type. In the first case, *exploitation*, a subcomponent of the argument's type is accessed and exploited, whereas in the second case, *introduction*, the selecting type is richer than the argument type and this last is wrapped with the type required by the function (cf. [Pustejovsky, 2006, Asher and Pustejovsky, 2006]). The reason why two kinds of coercion operation are proposed instead of one is that the information accessed in semantic composition can be differently embedded in a noun's semantics. In both cases, however, coercion is interpreted as a typing adjustment.

To begin, consider the standard selectional behavior of proposition-selecting verbs such as *believe*, *tell*, *know*, and *realize*. This can be seen in the range of data presented below.

(12) a. Mary believes [that the earth is flat].
b. John knows [that the earth is round].
c. John told Mary [that she is an idiot].
d. Mary realizes [that she is mistaken].

Using the typing convention introduced above, the argument structure for a verb such as *believe* would be given as shown in (13).

(13) $$\left[\begin{array}{l}\textbf{believe}\\ \text{ARGSTR} = \left[\begin{array}{l}\text{ARG1} = x : human\\ \text{ARG2} = y : info\end{array}\right]\end{array}\right]$$

While these are acceptable constructions, introspection suggests that these predicates also take non-proposition denoting expressions as arguments. For example, consider the sentences in (14) below.

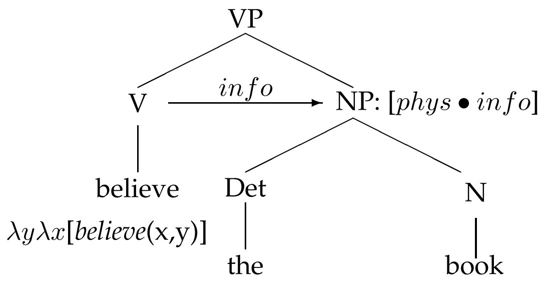
(14) a. Mary believed the book.
b. John told me a lie.
c. The man realized the truth.

Following [Pustejovsky, 1995], such expressions are licensed as propositional arguments to these verbs because they are "coerced" into the appropriate type by a rule of type exploitation. Specifically, as mentioned above, nouns such as *book* have double denotation. They are effectively "information containers", and can appear in contexts requiring both physical objects and information, as in (15).

(15) John memorized then burned the book.

The composition involved in a sentence like (14a) is illustrated below, where the informational component of the type structure for *book* is "exploited" to satisfy the type from the predicate.

(16)



This illustrates that predicates may have their selection preferences satisfied by exploiting the substructure associated with an argument. In this case, a propositional interpretation is construed from the type structure of the NP. Indeed, for each proposition-selecting verb in (12), there are well-formed constructions where an NP complement satisfies the propositional typing, as shown in (14).

3.2 Data Challenges Theory

In this section, we turn to naturally occurring (organic) data, equipped with the analytic framework from our first-level data analysis presented above. Using conventional and state-of-the-art tools in corpus analysis, we analyze the actual usage patterns for the predicates discussed above.

We analyze the set of complements for each verb by creating *lexical sets*,[2] a methodology first deployed in [Rumshisky et al., 2007] and used to examine selectional contexts for complex nominals.[3] We use the Sketch Engine [Kilgarriff et al., 2004] to examine the data from the British National Corpus [BNC, 1994]. The Sketch Engine is a lexicographic tool that lists salient collocates that co-occur with a given target word in the specified grammatical relation. The collocates are sorted by their association score with the target, which uses pointwise mutual information between the target and the collocate multiplied by the log of the pair frequency for a given grammatical relation. Additional corpus queries were performed using Manatee, a companion concordancing engine.[4]

In Tables 1, 2, and 3, we give the salient collocates for the verbs presented in the previous section, along with frequencies and association scores for each collocate. [5] We only list the complements that activate the relevant sense of the verb. For example, for the verb *realize*, we show the frequencies for propositional complements (e.g., *mistake, truth, importance, significance, implication, futility, danger, error*) and omit the complements activating the bring into being sense (e.g., *potential, ambition, dream, goal, hope, fear, ideal, expectations, vision, objective, plan*, etc.)

3.2.1 Case Study 1 (cont)

Our model predicts that the verbs *repair*, *fix*, and *mend* will select artifactual entities in direct object position, making implicit reference to the entity's Telic value. What we see in the actual data is that many of the complements do not refer to artifactual entities at all, such as: *damage*, *puncture*, *hernia*, *hole*, *crack*, *fault*, *problem*, *leak*, and *ravages* (cf. Table 1).

The problem that emerges from these data is that the same sense of each verb is being activated by semantically diverse lexical triggers, many

[2] Cf. [Hanks, 1996].

[3] See [Rumshisky and Batiukova, 2008] for more detail.

[4] See `http://www.textforge.cz/products`

[5] Sketch Engine word sketches for the BNC were manually edited to correct for misparses.

repair.v			fix.v			mend.v		
damage	107	42.66	pipe	9	11.83	fence	23	32.78
roof	16	20.27	gutter	4	11.45	shoe	10	19.01
fence	10	18.07	heating	5	9.66	puncture	4	18.91
gutter	5	15.87	car	19	9.43	clothes	11	18.68
ravages	4	15.76	alarm	5	9.13	net	8	18.01
hernia	4	15.61	bike	5	9.11	roof	8	16.99
car	23	15.39	problem	23	8.77	car	14	15.45
shoe	10	15.22	leak	3	8.58	way	20	14.26
leak	5	14.96	light	12	8.49	air-conditioning	2	12.71
building	17	14.02	boiler	3	7.96	damage	6	12.71
crack	6	13.99	roof	5	7.27	hole	5	11.38
wall	14	13.77	motorbike	2	7.19	bridge	4	9.68
fault	7	13.56	fault	4	6.91	heart	5	9.6
puncture	3	13.53	jeep	2	6.79	clock	3	9.45
pipe	7	12.89	door	11	6.65	chair	4	9.36
bridge	8	12.19	chain	4	5.48	wall	5	9.27
road	13	12.19	bulb	2	5.15	chain	3	8.3

Table 1: Direct object complements for the *repair*-verbs

of which are not artifactual objects. This raises the issue of the semantic relationship between these lexical items. These questions have been discussed previously in [Rumshisky, 2008] and [Pustejovsky and Jezek, 2008], and we turn to them with respect to the present case studies in Section 3.3.

3.2.2 Case Study 2 (cont)

For the next case study, our model predicts that NPs denoting information containers have the appropriate type structure to satisfy proposition-selecting predicates through type exploitation. That is, *the book* can denote a proposition in the sentence

(17) Mary believed the book.

Furthermore, we expect to see proposition-denoting NPs as complements as well. For example, *rumour* denotes a proposition in the sentence

(18) John doesn't believe the rumour.

We see from the data that there are many non-proposition-denoting NPs, varying from verb to verb. For example, for the verb *believe*, we have: *luck*, *eye, ear, tarot, woman, success*; for the verb *know*: *name, score, address, rules, trick*; for the verb *realize*: *futility, folly, threat, risk, cost*; for the verb *tell*: *history, ordeal, destination, suspicion, identity*, etc. The full list of complements, sorted by association score, is given in Tables 2 and 3. [6]

believe.v			know.v			realize.v		
luck	73	33.14	answer	389	35.17	mistake	15	20.02
ear	48	22.5	truth	219	30.92	extent	18	19.0
story	72	20.58	name	548	29.03	truth	15	18.7
word	95	19.02	whereabouts	37	24.64	importance	15	16.42
eye	74	15.19	secret	73	22.0	significance	11	16.11
hype	6	14.17	detail	142	17.77	implication	11	15.6
myth	12	14.07	story	141	17.48	futility	3	13.78
truth	19	13.31	meaning	78	16.58	value	17	13.28
lie	10	12.63	fact	159	16.28	danger	7	12.01
tale	13	12.61	reason	137	15.89	error	7	11.87
opposite	7	12.15	score	47	14.83	possibility	8	11.78
tarot	3	12.0	outcome	45	14.53	predicament	3	11.56
nonsense	7	11.6	saying	14	14.29	folly	3	10.09
propaganda	7	11.12	God	77	14.23	limitations	4	9.7
thing	47	9.12	username	7	14.02	strength	4	6.77
woman[7]	41	9.06	difference	105	13.98	need	6	6.07
fortune	8	8.82	feeling	79	13.75	threat	3	5.7
stupidity	3	8.57	word	162	13.74	benefit	4	5.31
rubbish	5	8.01	basics	10	13.53	problem	7	5.17
rumour	5	7.96	rules	99	13.03	advantage	3	5.04
evidence	19	7.81	address	42	12.74	difficulties	3	4.79
promise	7	7.78	password	10	12.4	effects	5	4.68
figures	21	7.78	identity	37	12.38	risk	3	4.68
forecast	5	7.49	joy	23	12.23	power	5	4.21
poll	7	7.48	trick	20	12.18	nature	3	3.7
gospel	4	7.45	place	171	11.88	fact	3	3.27
assurance	6	7.44	date	67	11.26	cost	3	2.94
success	14	7.35	extent	46	11.26			

Table 2: Direct object complements for the PROPOSITION/INFO-verbs

We clearly need to account for how these NPs satisfy the selectional conditions of the predicate, supposing our assumptions regarding the typing of the predicates are correct. Alternatively, we need to rethink the selectional specifications for each verb.

[6] For the verb *tell*, we give both direct object complements and NPs from ditransitive constructions, as identified by RASP parser [Briscoe and Carroll, 2002].

tell.v/direct object			tell.v/ditransitive obj2					
story	1286	52.0	secret	36	22.42	suspicion	4	5.62
truth	600	49.48	name	122	22.21	history	13	5.34
lie	254	45.67	detail	32	12.67	answer	9	5.33
tale	274	42.04	reason	37	11.06	direction	9	5.3
fib	18	30.84	gossip	6	10.4	dream	6	5.17
joke	94	28.85	ordeal	5	9.9	thought	10	5.08
untruth	8	19.08	gist	3	9.61	legend	3	4.92
anecdote	15	17.08	fact	34	9.5	age	13	4.7
difference	108	16.82	whereabouts	4	9.09	outcome	5	4.6
parable	8	12.75	trouble	9	6.98	symptom	4	4.32
fortune	24	12.57	plan	19	6.9	position	14	4.15
news	53	12.13	date	13	6.71	fate	3	4.08
			destination	4	6.54	identity	4	3.91

Table 3: Direct object and ditransitive obj2 complements for *tell*.

3.3 Theoretical Analysis of Structured Data

3.3.1 Case Study 1 (cont)

The first observation from analyzing organic data associated with the selectional behavior of verbs like *fix*, *repair* and *mend* is that there are, in fact, two major selectional clusters, not one. One indeed involves the artifactual entities as predicted by our theoretical assumptions. The other, however, refers to a negative stative or situational description of the artifactual under discussion. Further, we observed that this latter cluster divides systematically into two classes, one a general negative situation, and the other referring to the condition of the artifact, as can be seen in lexical sets in (19), (20), and (21). [8]

(19) *fix.v*
object
a. ARTIFACTUAL: pipe, car, alarm, bike, roof, boiler, lock, engine; heart; light, door, bulb
b. NEGATIVE STATE (condition on the artifact): leak, drip
c. NEGATIVE STATE (general situation): problem, fault

(20) *repair.v*
object
a. ARTIFACTUAL: roof, fence, gutter, car, shoe, fencing, building, wall, pipe, bridge, road; hernia, ligament
b. NEGATIVE STATE (condition on the artifact): damage, ravages, leak, crack, puncture, defect, fracture, pothole, injury
c. NEGATIVE STATE (general situation): rift, problem, fault

[8] Semicolon is used to separate semantically diverse elements of each lexical set.

(21) *mend.v*
object
a. ARTIFACTUAL: fence, shoe, clothes, roof, car, air-conditioning, bridge clock, chair, wall, stocking, chain, boat, road, pipe
b. ARTIFACTUAL (extended or metaphoric uses): matter, situation; relationship, marriage, relations
c. NEGATIVE STATE (condition on the artifact): puncture, damage, hole, tear

Assuming that these are all instances of the same sense for each of the verbs, how do we incorporate these observations back into the selectional properties of the verb? First, as mentioned above, there appear to be two negative states selected in many cases:

(22) a. GENERAL NEGATIVE SITUATION: "fix the problem"
b. CONDITIONS OF THE ARTIFACT: "hole in the wall", "dent in the car".

What do these clusters have in common? Does the verb select for either a negative situation or an artifact? The answer is: basically, the verbs select for a negative state of an artifactual.

When the negative relational state is realized, it can either take an artifactual as its object, or leave it implicitly assumed:

(23) a. *repair the puncture / leak*
b. *repair the puncture in the hose / leak in the faucet*

When the artifactual is realized, the negative state is left implicit by default.

(24) a. *repair the hose / faucet*
b. *repair the (puncture in) the hose / (leak in) the faucet*

This suggests that the theoretical description of the selectional properties for the verb *repair* needs modification to reflect behavior witnessed from the organic data. This can be accomplished by positing the negative state as the selected argument of a verb such as *repair*, and the artifactual posited as a *default argument*.

(25)
$$\begin{bmatrix} \textbf{repair} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = x : human \\ \text{ARG2} = y : neg\_state(z) \\ \text{D-ARG1} = z : phys \otimes_{Telic} \alpha \end{bmatrix} \end{bmatrix}$$

This has the effect of explaining the lexical set distribution: when the noun denotes a negative state, there is an implicit (default) artifactual quantified in the context. When the artifactual is realized, the negative state interpretation is present in a type of coercion (introduction). Hence, both patterns are accounted for by the lexical structure for the verb along with compositional principles allowing for coercion.

3.3.2 Case Study 2 (cont)

From examination of the data on NP-complements to proposition-selecting predicates, we see that type coercions, when they exist, are distributed in very different ways for each verb. Theoretically, this means that the licensing conditions for type coercion must be distinct in each of these cases. Given the theoretical fragment we presented in Section 3.1.2, however, there are no mechanisms for explaining this distribution.

In order to understand this behavior better, let us examine the non-coerced complementation patterns of these verbs in corpora. Several subclasses of clausal complements are attested in the BNC for each of these verbs. Namely, we identify the following three complement types:

(26) a. FACTIVE: *know*, *realize*
b. PROPOSITION: *believe*, *tell*
c. INDIRECT QUESTION: *know*, *tell*

We have already encountered the syntactic behavior of propositions in (12). The class of "factives" includes verbs that presuppose the situation denoted by the complement. For example, in (27), the situation denoted by the complement is presupposed as fact.

(27) a. John realized [that he made a mistake].
b. Mary knows [that she won].

The class of "Indirect questions" includes verbs selecting a *wh*-construction that looks like a question, but in fact denotes a value. For example, the verb *know* allows this construction, as does *tell*:

(28) a. Mary knows [what time it is].
b. John knows [how old she is].

(29) a. Mary told John [where she lives].
b. John told me [how old he is].

In order to account for this data, the model must allow each verb to carry a more specific encoding of its complement's type than we had initially assumed, except for the verb *believe*. This suggests the revised argument structures [9] below.

(30) **believe**(ARG1:$human$, ARG2:$prop$)

(31) a. **tell**(ARG1:$human$, ARG2:$info$)
b. **tell**(ARG1:$human$, ARG2:$Ind\_Question$)

(32) a. **know**(ARG1:$human$, ARG2:$factive$)
b. **know**(ARG1:$human$, ARG2:$Ind\_Question$)

(33) **realize**(ARG1:$human$, ARG2:$factive$)

The question is whether these verbs have the same semantic selectional behavior when occurring with NPs as they do with clausal complements. Consider first when an NP can be interpreted as an indirect question. What we see in the corpus is that one set of arguments for the verbs *know* (and *tell*) includes nominals that denote the value of something interpreted as a varying attribute; that is, they can take on or assume the interpretation of an indirect question in the right context. For example, the noun *age* is an attribute of an object with different values, and the noun *time* in this same context can be interpreted as an indirect question.

(34) a. Mary knows the time.
b. John knows her age.

(35) a. Mary told John her address.
b. John told me his age.

This NP construction is usually referred to a "concealed questions" structure. The lexical sets for the verbs *tell* and *know*, organized by most probable semantic type, are shown in (36) and (37) below. The BNC data in these lexical sets was collected using the Sketch Engine, and manually sorted according to the complement type.

(36) *tell.v*
object
a. PROPOSITION: story, truth, lie, tale, joke, anecdote, parable, news, suspicion, secret, tale, details, gossip, fact, legend; dream, thoughts
b. INDIRECT QUESTION: name, whereabouts, destination, age, direction, answer, identity, reason, position, plan, symptoms; outcome, trouble

[9] The feature structure notation is simplified for readability and space considerations.

(37) *know.v*
object
a. FACTIVE: truth, secret, details, story, meaning, fact, reason, outcome, saying
b. INDIRECT QUESTION: answer, score, whereabouts, address, username, password, name; feeling, difference

With the verb *realize*, the data show that NPs complements can also assume a factive interpretation:

(38) John realized his mistake.

But what is interesting is that the majority of the nominals are abstract relational nouns, such as *importance*, *significance*, *futility*, and so forth, as illustrated below.

(39) *realize.v*
object
FACTIVE: importance, significance, extent, implication, futility, value, error, predicament

For the verb *believe*, all nominals are coerced to an interpretation of a proposition, but through different strategies. Those nominals in (40a) either directly denote propositions (e.g., *lie*, *nonsense*) or are complex types that have an information component which can interpreted propositionally (e.g., *bible*, *polls*). The sources in (40b) are construed as denoting a proposition produced by (e.g., *woman*), or coming through (e.g., *ear*) the named source. Finally, the last set is licensed by negative polarity context, and is a state or event; e.g., "He couldn't believe his luck.").

(40) *believe.v*
object
a. PROPOSITION: lie, tale, nonsense, myth, opposite, truth, propaganda, gospel
b. SOURCE: woman, government, bible, polls, military; ear, eye
c. EVENT/STATE: luck, stupidity, hype, success

Note also that the prediction that selectional specifications of *believe* as an information-selecting predicate could be satisfied by any information nominal is not borne out. For instance, the informational component of a complex type $phys \bullet info$ does not seem to encourage the interpretation appropriate for a complement of *believe*. While some nouns of $phys \bullet info$ type, such as *letter*, do accept this interpretation, it is so infrequent that it is not attested in roughly 33,000 of occurrences of *believe* in the BNC. Other nouns of $phys \bullet info$ type, such as *novel*, do not seem to be capable of this interpretation altogether.

This also suggests that different information-selecting predicates in fact require different propositional structure from the complements. For example, *believe* requires the informational noun to allow either a single message interpretation (e.g. *believe the nonsense*) or a source interpretation (e.g. *believe the political blogs*).

This necessity for refinement of selectional specifications is also apparent for other information-selecting predicates, for example, for *write*. In the classic GL interpretation, this verb selects for the artifacts of $phys \bullet info$ type with Agentive "write" and Telic "read" – that is, they select for objects that are produced by writing and whose purpose is to be read. But consider the nouns in (41) which clearly match this specification, and yet differ in their ability to satisfy the corresponding selectional requirements.

(41) a. John wrote a novel.
b. ?John wrote a dictionary.
(but cf. "You have to love a lexicographer who had the courage, interest, and patience to write an entire dictionary by himself.")
c. ?John wrote a newspaper.
(but cf. "Sixth-grade pupils wrote a newspaper for their parents describing their experiences in different curriculum areas in the classroom."

While (41a) is acceptable without qualification, both (41b) and (41c) require a bit of context to modulate the composition, enhancing the "naturalness" of the expression, in Sinclair's sense [Sinclair, 1984]. So in fact a more refined specification is needed to explain combinatorial behavior of these nouns, one perhaps taking into account the exact manner in which information carried by each artifact is produced.

4 Concluding Remarks

In this paper, we have examined the contributing roles both corpus-based and model-based linguistics play in constructing an adequate characterization of language usage. By its very design, Generative Lexicon aims to explain the contextual modulations of word meanings in actual data. Therefore, the distributional profile presented by large corpora is not a tension so much as a necessary component to a healthy investigation of the phenomenon, namely, the infinite richness of language. While the generative notion of the *ideal speaker/hearer* of language is a powerful notion, it is empty without application and revision through data. As Sinclair has aptly stated:

> Starved of adequate data, linguistics languished. . . . It became fashionable to look inwards to the mind rather than outwards to society. [Sinclair, 1991]

References

[BNC, 1994] (1994). *The British National Corpus*. The BNC Consortium.

[Asher and Pustejovsky, 2006] Asher, N. and Pustejovsky, J. (2006). A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6:1–38.

[Bouillon, 1997] Bouillon, P. (1997). *Polymorphie et semantique lexical: le case des adjectifs*. PhD dissertation, Paris VII, Paris.

[Briscoe and Carroll, 2002] Briscoe, T. and Carroll, J. (2002). Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002*, pages 1499–1504.

[Church and Hanks, 1990] Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.

[Fillmore, 1991] Fillmore, C. (1991). 'corpus linguistics' vs. 'computer-aided arimchair linguistics'. Berlin. Mouton de Gruyter.

[Firth, 1957] Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Philological Society.

[Hanks, 1994] Hanks, P. (1994). Linguistic norms and pragmatic explanations, or why lexicographers need prototype theory and vice versa. In Kiefer, F., Kiss, G., and Pajzs, J., editors, *Papers in Computational Lexicography: Complex '94*. Research Institute for Linguistics, Hungarian Academy of Sciences.

[Hanks, 1996] Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1).

[Hornby, 1954] Hornby, A. S. (1954). *A Guide to Patterns and Usage in English*. Oxford University Press.

[Kilgarriff et al., 2004] Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

[Moravcsik, 1975] Moravcsik, J. M. (1975). Aitia as generative factor in aristotle's philosophy. *Dialogue*, 14:622–636.

[Pustejovsky, 1991] Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4).

[Pustejovsky, 1995] Pustejovsky, J. (1995). *Generative Lexicon*. Cambridge (Mass.): MIT Press.

[Pustejovsky, 2001] Pustejovsky, J. (2001). Type construction and the logic of concepts. In *The Syntax of Word Meaning*. Cambridge University Press, Cambridge.

[Pustejovsky, 2006] Pustejovsky, J. (2006). Type theory and lexical decomposition. *Journal of Cognitive Science*, 6:39–76.

[Pustejovsky and Jezek, 2008] Pustejovsky, J. and Jezek, E. (2008). Semantic coercion in language: Beyond distributional analysis. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*. forthcoming.

[Rumshisky, 2008] Rumshisky, A. (2008). Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*. forthcoming.

[Rumshisky and Batiukova, 2008] Rumshisky, A. and Batiukova, O. (2008). Polysemy in verbs: systematic relations between senses and their effect on annotation. In *COLING Workshop on Human Judgement in Computational Linguistics (HJCL-2008)*, Manchester, England. submitted.

[Rumshisky et al., 2007] Rumshisky, A., Grinberg, V. A., and Pustejovsky, J. (2007). Detecting Selectional Behavior of Complex Types in Text. In Bouillon, P., Danlos, L., and Kanzaki, K., editors, *Fourth International Workshop on Generative Approaches to the Lexicon*, Paris, France.

[Sinclair, 1966] Sinclair, J. (1966). Beginning the study of lexis. In Bazell, C. E., Catford, J. C., Halliday, M. A. K., and Robins, R. H., editors, *In Memory of J. R. Firth*. Longman.

[Sinclair, 1984] Sinclair, J. (1984). Naturalness in language. In J.Aarts and Meijs, W., editors, *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Rodopi.

[Sinclair, 1987] Sinclair, J. (1987). The nature of the evidence. In Sinclair, J., editor, *Looking Up: An Account of COBUILD Project in Lexical Computing*. London: Collins.

[Sinclair, 1991] Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

[Sinclair, 2004] Sinclair, J. (2004). *Trust the Text*. Routledge.