B978-0-32-385197-8.00016-7, 00009

CHAPTER

9

Conceptual spaces and scientific data models

OUTLINE

9.1 Introduction 233

9.2 Verb-centric grammars and information-delta paths 236
- 9.2.1 *The emergent syntax/semantics interface* 243

9.3 Conceptual and thematic roles 244
- 9.3.1 *Disjoint conceptual spaces* 247
- 9.3.2 *Conceptual spaces and scientific data* 255

9.4 Delta roles and conceptual space markup language 257
- 9.4.1 *Information delta and data modeling* 257
- 9.4.2 *The artificiality of data semantics* 260

9.5 Conclusion: Toward a scientific data semantics 261
- 9.5.1 *Research data and data integration* 263
- 9.5.2 *Toward a procedural conceptual-space semantics* 264

References 266

9.1 Introduction

In the two decades since conceptual space theory was first proposed (in a mostly cognitive-linguistic setting), researchers working in a more strictly computational framework have adopted, extended, or formalized Gärdenfors's approach in several different ways. Some have focused on potential AI applications; others on data-modeling and the systematization of scientific data descriptions; and some computational linguists (including via Quantum NLP) have proposed developing a computationally tractable semantics rooted in conceptual spaces.

Work such as Coecke *et al.* [14], which we emphasized in Chapter 6, intimates that conceptual spaces offer a path toward resolving the "grounding problem," wherein mechanical systems can never truly have semantics in the first place—there is never a deliberate or conscious correlation of system state to external

NEUSTEIN, 978-0-323-85197-8

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:55 • p. 1
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

objects/situations—but merely a simulacrum of semantics engineered to force "internal" and "external" state into a rough alignment, with the effect of endowing AI machines with certain behavioral tendencies that we associate with language-use (and human intelligence in general). Investigations related to "solving" the grounding problem have been among the primary motivations for incorporating conceptual spaces in an AI context (*see* [23], [24], [19], [17], [41], [3], or [4] for good examples). On the other hand, work such as Coecke *et al.* also suggests conceptual space applications to the issue of *construction semantics*, or clarifying how meaning-compositions (which we can also characterize as *concept blends*, so long as there is no confusion with other uses of this phrase, notably Gilles Fauconnier's and Mark Turner as in [20], [21]), such as "dark red," "red square," "wedding ring," "fake diamond," and so forth, signify a compound idea that structurally integrates its parts.

The possibility of finding quantitative dimensions within concept-extensions serves a useful purpose with respect both to concept-grounding and concept-blending. With respect to grounding, dimensional quantification gives us an idea of how concepts may take shape from recurring exposure to certain perceptual or situational patterns, as well as concrete models of concept-learning that could be simulated via neural networks. With respect to concept-blending, dimensional models are obviously applicable to constructions such as *dark red* and *red square*, where concept-blends present intersections or unions of dimensions in straightforward ways (darker hues of red and red-hues plus square-shapes, respectively).

Classifying constructions according to dimensional interactions (e.g., intersection, union, or contrastive) presents a useful semantic outline that applies to many forms of constructions.[1] As we argued in Chapter 6, an equally important construction genre is more exactly modeled by considering concepts playing different roles in larger situations, in which case the *union* between multiple concepts is characterized by concepts being indexed by roles with relatively little inter-dimensional interaction. With that caveat, though, conceptual space models seem to have a solid place in the general system of situational semantics (viz., semantics of situations in general, including but not restricted to information-theoretic situation theory associated with Jon Barwise, John Perry, and Joe Seligman).

The form of syntax/semantics interface pursued by Coecke *et al.* which we reviewed in Chapter 6 has been developed by the same research team in relation to several different semantic paradigms, including statistical word-use correlations, which do not purport to offer a bonafide semantics at all (an AI agent which infers properties such as topic-relevance based on statistical correlations across document corpora, without attempting to reason about the "meaning" of any words apart from noting how likely or unlikely they are to appear in a given context, is not "grounding" symbols at all, but rather using probabilities to approximate the behavior of a cognitive system). Coecke *et al.* therefore explicitly state that one motivation for pursuing conceptual spaces as the foundation for their semantic layer is to provide a more cognitive semantics to pair with a grammar based on hypergraph categories. In this sense, their research fits within the "grounding" topic in terms of the methodological rationale for incorporating conceptual spaces. In substance, however, the models they actually formulate—which can

[1] *See* for example [1, page 9]; contrast classes are modifiers which tend to restrict the scope of a conceptual space down to a subregion, as in examples like the "small" of *small planet*, but do not have dimensional structures themselves, because *small* for example can imply many different size-ranges depending on the scale of object to which it is applied (*small planet* vs. *small molecule*).

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:55 • p. 2
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

be glossed as employing a category-theoretic framework to formalize the basic idea that syntactic constructions represented via hypergraph categories can engender semantic constructions analyzed via concept-blending (recall the specific sense of *engender* we adopted in Chapter 6)—are thematically aligned more with construction semantics.

In the context of natural-language theories, to be sure, grounding and constructions can be philosophically interconnected: cognitive construals of integrated situations obviously emerge from cognitive-perceptual discernment of phenomenologically fundamental concepts (so Gregor Strle, for instance, considers the possible role of conceptual spaces as "a mediating level between symbolic and subconceptual representations," albeit less crisp in actual cognition than in the artificially simplified terms of the theory: "methods and models ... are ... explanatory tools, i.e. instruments used to explore, simulate or explain particular aspects of cognition, not mechanisms of human mind." [49, page 18]) As the last quote intimates, neat quantitative models of concept-acquisition/emergence need not be read as realistic simulations of cognitive processes; instead, this like any other theory presents an unrealistically simplified but explanatorily useful model of the actual systems it seeks to clarify. The real conceptual "spaces" in the mind may have many more dimensions, with less obvious numerical encodings, than colors or spatial configurations.

If we accept the simplifying gestures of axis-like concept dimensions as an imprecise but heuristic gloss on cognitive explanantia, then we can imagine *something like* an emergence of concepts by dynamic clustering of similar experiences (and resolving borderline cases intermediate between related concepts) being among the key driving forces of cognitive symbol-grounding. Likewise we can picture convolution between different concepts' dimensions as defining the contours of the hybrid concepts resulting from semantic blends (and hence semantic constructions, engendered by syntactic concatenations between concept-words). The cognitive dynamics and dimensional structures in both cases are similar (governed by the same conceptual space models), which suggests—even if we assume that real-world cognitive processing is much subtler than such toy models articulate—that there is some cognitive resonance, some redeployment of a common neurological architecture, in the primitive concept-grounding phenomenon as well as the more mentally sophisticated construction-semantics hermeneutic.

Arguably, however, issues of grounding and constructions are more strictly separate when considering the "semantics" of formal data structures, or, in practice, developing working semantic models for resources such as scientific data sets. Certain themes which are central to conceptual space theory clearly have direct applications to computational data modeling, such as dimension/domain structures and region-based classification within and among concepts (for prototypes, concept extensions, borderline-case disambiguation, and so forth). However, rigorous dimensional modeling and geometric clustering are hardly unique to conceptual space theory, and when considering applications of this specific theory to formal data semantics, we should attempt to isolate philosophical features endemic to the conceptual space lineage in itself (as opposed to broadly applicable statistical paradigms). Conceptual space theory's manifestations in the specific natural-language concerns of symbol-grounding and construction-semantics do not necessarily carry over to formal data semantics, or to such practical applications as query-engine implementation or data-set annotation semantics.

In this chapter, by contrast, we will consider the degree to which different use-cases or adoptions of conceptual space theory in a computational setting *can* be synthesized and integrated, focusing attention on practical application within the general scope of *data* seman-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

tics. Part of this analysis involves generalizing Coecke *et al.*'s model of the syntax/semantics interface to, potentially, the realm of formal/programming languages, continuing Chapter 6's "syntagmatic graph" model. That model provides a framework for codifying the relation between formal-language *syntax* with *semantics* in terms of a notion of "information content" (which we will make more precise). We will argue that this framework presents a foundation for integrating the data-modeling and scientific-computing conceptual space use-cases, and in particular that the relevant notions of information content provide an additional formalization of notions central to the original Gärdenfors theory (such as dimensional separability and domain correlation).

9.2 Verb-centric grammars and information-delta paths

Our syntagmatic graph approach assumes a specific reading of Coecke *et al.*'s syntax-semantic interface theory, albeit (we would argue) one which seems well-motivated by their specific text (and moreover is sufficiently general as to be applicable in a variety of contexts): paths in syntagmatic graphs (or analogous syntax-representing structure/spaces) map to "semantic" paths *characterized by an increase in information content*. Each step along a syntagmatic path is mirrored by a correlative step—whose nature is fixed by the underlying syntactic neighborhood (this is precisely the dynamics of grammatical principles; a language's syntactic system should carry enough determinateness that syntagmatic steps, e.g., between word-pairs, impute semantic modifications in a well-determined way)—a semantic "space" culminating in a sentence's overall meaning (which is the end-point of a semantic path, corresponding to a sentence's root verb as the end-point of a syntactic path).[2]

Semantic paths should follow syntagmatic paths fairly predictably, that is, for a given syntagmatic step, there should not be a multitude of noticeably different semantic steps, which are all possible correlates of the syntagmatic step according to the language's syntax-semantics interface (semantic steps, which are underdetermined, would result in large-scale sentence-level ambiguity). On the other hand, "points" and "steps" in semantic space should not be pictured in too crisp a manner; the full meaning of a sentence only emerges in its entirety, and there may be degrees of detail that remain unresolved while the sentence is unfolding (for instance, postcedent anaphoric resolution, as in *they* to *England* in *although they were playing in London, England lost the 2020 final*). Syntagmatic steps should "compel" paths in semantic space, but not too rigidly; this interplay between syntax and semantics creates a dynamic environment, where syntactic norms (and lexical specificity) evolve toward the proper balance of rigor and flexibility. Or at least, this is a plausible working model of the syntax/semantics interface.

Assuming some form of this hypothesis is in effect, we can picture "paths in semantic space" as the *image*, under certain (not-fully crisp) mappings between syntactic and semantic systems, of syntagmatic paths. In our proposed syntagmatic graph model, syntagmatic paths are characterized by the structural tendencies of the underlying graph (in particular they are constrained to lead toward verbs/procedure-nodes except via specific "hinge" nodes or edges), and semantic paths via the general stipulation that movement along the path corresponds with amplification of information content. The overall syntax-semantics interface then ensures (and evolves to satisfy the need for) these two path-

[2]Here we continue the analysis developed in the first section of Chapter 6; we defer to that chapter for definitions of nonstandard terms.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

tendencies to operate in parallel. Syntagmatic steps toward verbs/procedures should correlate with amplificatory steps in a semantic space permeated by a measure of information content, and vice-versa.

This picture, though intuitively compelling (we suggest), leaves quite open-ended any specification of what the parameters of "semantic space" actually are. Coecke *et al.* address (what appears to be) an analogous issue by examining the interpenetration of conceptual spaces. It is helpful (even if just as a mental figure) to picture "semantic space" as a space with points and regions, and conceptual space theory legitimates such a picture, because according to this theory many concepts can indeed be given geometric form, in terms of domains and dimensions, which possess at least some quantitative characteristics (e.g., the optical mathematics of colors, or the mereotopological interpretations of spatial terms, such as *near*, *over*, *across*, *around*, *inside*, *alongside*, etc.). Coecke *et al.* try to generalize this picture by arguing that (albeit extended to more abstract forms of quantifiability, e.g., the similarity of an object to a prototype) all concepts can be expressed, via suitable choices of dimension, in *some* geometric terms, so we have a basis for modeling concept-combinations in general as unions and intersections between quantifiable spaces. Such generalized spaces therefore provide us with territory on which to define the general notion of "semantic paths," which (even if different terms are used) are the central facets of analyses explaining how syntactic form engenders semantic meaning. Once we have a rigorous model of semantic space, we can develop a theory of semantic paths mirroring (what we are calling) syntagmatic paths.

Since we have argued that paths follow *directions of increase* in *information content*, we will use the generic term "information delta" to designate the degree and qualitative facets of such amplification: *why* does information elevate along a path and *how much*. Clarifying information delta along "paradigmatic paths," or paths in *semantic* space, we claim, furnishes a starting point for explaining how *syntagmatic paths* engender (in the sense of mostly unambiguously determining) semantic constructions. This dynamic can be observed in both natural and programming languages. We will devote the majority of discussion to the latter case, but we will motivate our analysis with a brief review of the natural-language case. This provides us an opportunity to summarize a natural-language theory for syntagmatic graphs that was alluded to (but postponed) in Chapter 6.

Conventional conceptual space theory usually talks about "regions" more than "paths." For example, the "blend" of two concepts represents the intersection of regions, where the two concepts share similar dimensions (e.g. *dark red* or *shiny red* represent combinations of color and/or optical properties, which generally share dimensions and therefore yield mixtures that reduce the scope of applicability; *dark red* is more specific than either *red* alone or *dark* alone) as well as the union of dimensions, which are not shared (e.g., *red square* defines a concept whose dimensional boundaries include both spatial-configuration and color-optical domains). Implicitly, Coecke *et al.*'s insight is that anywhere we have enough space-like structure (or, say, topos-like structure, since in their context we are working in a categorial framework) to define *regions* (as in the Gärdenfors "concepts are convex regions"), then we also have enough structure to consider *paths*, and whereas concepts are *regions* in that "space," paths are *meanings* in the space, derived from language-like artifacts (structures with a grammar), in the sense that the meaning of a whole corresponds with a path end-point, and the meanings of parts correspond to path segments.

A general notion of "semantic space" accordingly has to serve several theoretical goals: accounts of semantic *regions* should buttress theories of concepts, while *paths* should be analytically associated with "meanings" *under the assumption* of a syntax-semantic interface, where

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:55 • p. 5
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

syntactic rules govern how syntagmatic steps compel semantic steps. Tying these two ideas together, paths in semantic space tend to integrate multiple concepts (correlative to syntagmatic paths visiting tokens that encode those concepts), so *red square* iconifies a step defined via syntax (adjective modifying noun), a step in some semantic space (arriving at the idea of a red square, which can then be plugged in to more general semantic context), and a blend of two concepts (the merged idea carries concepts of both *redness* and *squareness* into the contexts where it is used). Conceptual space theory has tended to gravitate toward blending scenarios, where concept-juxtapositions *narrow the scope* of each concept: *dark red* is narrower than *red* alone, because "dark" filters our some red shades. Likewise *red square* is also a kind of narrowing—even though *red* and *square* do not restrict one another within common dimensions (although their dimensions are not entirely separate, since both involve a minimal sense of spatial *location* even if not spatial *region*)—because *red square* can only be predicated of something with both spatial configuration and optical properties (not *red light*, say). It may be intuitively natural to highlight such "narrowing" effects of concept-combinations due to the overarching theoretical program, wherein semantic paths converge to precise meanings: if each successive step combines two concepts, then paths which are chains of such steps evince greater conceptual precision. On that intuition, the examples of conceptual spaces generating narrowing effects from concept-juxtapositions would seem like especially important case studies for the general project of using conceptual spaces to define "semantic" space in general.

This chapter will argue, however, that concept-blend analyses are of only limited value for either natural or formal languages, and will instead focus on issues such as *conceptual roles* and *dimensional analysis*, which will integrate the data-modeling applications of conceptual space theory with the linguistics-oriented themes of the last few paragraphs. In particular, we will continue the discussion of role-indexed multipart relations initiated toward the end of Chapter 6.

We should be careful not to take analogies between natural and programming languages too far: many correspondences between the two forms of languages are valid but superficial, so they do not particularly engender new ideas or technical foundations either for natural linguistics or for programming language design or implementation that would actually advance those disciplines. Nonetheless, even fairly superficial or vaguely specified correlations between theories (or models, perspectives on, and so forth) formal and natural languages can be useful "intuition primes," suggesting new lines of research, which could plausibly yield more rigorous results than the underspecified initial intuitions might intimate. For example, the semantics adopted by Quantum NLP—at least as pursued by the Oxford Quantum Group—arguably bears only a superficial resemblance to conceptual space theory as formulated by Gärdenfors and subsequent linguists, who were working primarily in a cognitive linguistic and largely humanities/philosophical context. At least, once we consider the kinds of semantics that may actually be modeled according to Quantum logic, i.e., that would currently run on the Oxford group's quantum computing environments. However, investigating the philosophical resonance between conceptual spaces and the form of semantics endemic to Quantum NLP helps to illustrate—at least if we consider work such as Coecke *et al.* to be persuasive overviews of the relevant technical issues—that conventional formal-language "semantics" (leaving aside questions of "natural" semantics aligned with human pragmatics and conceptualizations) has its own theoretical limitations.

This chapter will consider one specific model or perspective, which cuts across both formal and natural languages, although with the above provisos in effect: we should be wary of plausi-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

ble but largely vacuous correlations that might be identified between these two genre of languages, and consider any perspectival arguments or technical models to have merit only if they appear to lead in practically useful directions for either natural linguistics or software language engineering and related fields. We need to try to find ideas of theoretical substance from the larger space of basically trivial summarial notions, which are largely self-evident when considering language in general (both formal and natural) and syntax/semantics.

It is obvious, and as such not especially useful to incorporate into a rigorous theory, to note the general point that grammar governs which words are linked with which other words, and that word-pairs have semantic interpretations wherein the pair has a conceptual detail, modification, or specificity, which is lacking for either word in isolation. For instance, *beautiful diamond*, *antique diamond*, *fake diamond*, *expensive diamond*, *her diamond*, *some diamond*, and *the diamond* all ground, modify, and/or qualify the generic concept "diamond" in some fashion. This is an example of how word-pairs "compose meanings."

Understanding a sentence—at least according to a wide range of linguistic schools of thought, which tend to leverage this intuition in different ways—is in large part a question of identifying which words to pair with which other words so as to compose meanings "the right way": a sentence is a "composition" built up from word-pairs which are themselves compositions fusing two "meanings." The crux of sentence structure is identifying which words are directly connected to other words to form the building-block "compositions," which are fused together in the larger sentence, e.g., in *the brilliant diamond on her sister's expensive diamond wedding ring* note that *brilliant* is composed with *diamond*, as is *the*; *expensive* is composed with *ring*, as is *wedding* and the second *diamond*; and *her* is composed with *sister*. In natural language, words can pair up according to rules that are not evident in the surface language; for instance, it is not the case that an adjective always immediately precedes a noun (in the above sentence *expensive* modifies *ring* although there are two other words between them; and the possessive links *sister* to *ring*, even though the two words are far apart). The fact that meaning-composition may involve words, which are superficially distant, and where there may not be explicit markers (such as morphological cues) pointing to correct word-pairings, suggests that syntactic rules serve the specific purpose of establishing which word-pairs are compositionally relevant for any given sentence.[3]

This picture has theoretical ramifications that might not be obvious if we only consider the obvious point that meaning is compositional; that the meaning of a sentence is somehow a product of its component parts. The more specific argument is that compositionality has a certain internal structure, and that we can model composition in terms of *sequences of word-pairs*, where the role of grammar is first to identify which word-pairs are in effect, and second to "rank" the pairs as logically prior or posterior. An open question is how to model the "space" that emerges if we see the ordering of word-pairs as forming a "path." One criterion of this space should be that paths culminate in something like *propositions* (or "complete ideas"), and also that there is some ambient notion of "information content," such that paths progress *in*

[3]Moreover, because word-distance is not a definitive criterion for the validity or non-validity of sentence-parses in word-pair sets, there is a vast collection of possible parses, which are plausible based solely on type-reduction criteria (or whatever combinatory or phrase-structural constraint play roles akin to type-reduction in a pregroup-compatible grammar). This is one reason why Quantum methods may be feasible and yield superior performance to classical computations: the initial goal of NLP is to derive a parse-graph, which maximizes "coherence" amongst many parses that are at least *somewhat* coherent. We can, potentially, define coherence in terms of word-pairing by saying that a given (potential) word-pair has greater coherence if the information content yielded by the pair has greater increase on that of each word in isolation than alternative potential pairs.

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 7
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

the direction of greater information content on their way to the "sentence" terminus. A further implication is that sentences have *more* information content (however this is modeled) than their component parts.

This principle is obliquely raised by Coecke *et al.* in the goal of developing a linguistic type system such that "type reductions in the grammar category [map] onto algorithms for composing meanings." Insofar as the *purpose* of grammar, or at least one of its most essential roles, is to govern how addressees decompose a sentence into word-pairs (which semantically "compose meanings"), natural language grammar presumably evolves under pressures to play this role effectively. Parts of speech, for example, can be differentiated in terms of their corresponding roles in word-pair formations; adjectives always modify nouns, for instance, whereas nouns get attached to verbs as subjects and/or objects.

Languages therefore tend to separate different parts of speech through lexical convention (most words' primary meanings are associated with one part of speech in particular) and/or morphological markings (such as modifications, which transform lexemes between parts of speech). In English, for example, the *-ly* suffix converts adjectives to adverbs (*quick* becomes *quickly*). Subtler morphological cues are evinced in case-markings or declensions, which register a noun as subject or object, say (nominative or accusative case, for languages that, more so than English, feature substantial declension markings) or as connected to a verb in some other manner governed by a language's case system: locative, instrumentive, dative, benefactive, and so forth. This general principle of parts-of-speech corresponds, in NLP and computational linguistics, to the idea that lexemes have *types*, and that rules governing when words can be paired up can be modeled via formal type systems (hence the idea of "type reductions"). In short, what we as language users experience as the natural synergy between corresponding parts of speech, which engender a semantically meaningful pairing (like *expensive ring*) corresponds formally to a type-reduction rule, whereby the collision between two types yields a third type.

This type-reduction phenomenon exists on the scale of individual word-pairs, but a central theory of formal linguistics is that type-reductions can be chained iteratively, which captures the semantic notion that meaning-compositions build up to a "complete idea." For instance, after composing *diamond* and *ring*, we can add a further composition (e.g., *expensive*), or a grounding (*the diamond ring*) or possessive (*her diamond ring*), or supply a verb for which the present meaning supplies a subject (*her diamond ring was an anniversary present*). The *pattern* of meaning-composition is governed by grammar, which provides the essential detail of clarifying the proper *order* of composition, which in turn can subtly (or radically) alter a sentence's overall meaning. For instance, *her anniversary present was a diamond ring* connotes something somewhat different, especially in context, than the inverted sentence (*her diamond ring was an anniversary present*); still more noticeably, *she divorced him* is different than *he divorced her*, *she gave him that ring* different from *he gave her that ring*, and so forth. If considered on the basis of their lexical senses alone—i.e., if we did not have syntactic rules as a further source of "information" about speaker intentions—words in a typical sentence can be paired up in many different ways. Syntactic rules therefore need to be sufficiently rigorous that most sentences are decomposable into constituent meaning-compositions (i.e., word-pairs) without undue ambiguity, because they are a key source for clarifying which words are intended to be connected (excluding only word-pairs that make no conceptual sense leaves too many options still available). For this to work, syntax has to be fairly rigorous and rule-bound, arguably implying that it may be

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

summarized via a formal machinery, such as computational-linguistic type systems.[4]

Coecke's *et al.*'s theory therefore leverages the well-established idea that meaning is compositional, and that this compositionality can be structurally divided into individual word-pairs, where syntactic rules govern which word-pairs are semantically in effect for a given sentence. Their use of category theory to define notions, such as "pregroup grammars," largely plays the role of formalizing conditions on how words "pair up": constraints that may be expressed through mathematical formalisms (such as category theory) act as filters—or, more precisely, somehow model cognitive activities that act as filters—which select *some* word-pairs as in effect for a given sentence, screening out others that are lexically plausible, but semantically inaccurate in context (e.g., in *expensive diamond ring* it is, explicitly, the *ring* which is described as expensive, not necessarily implying that there is a single expensive diamond on that ring).

In addition, as well as *filtering* word-pairs, grammatic rules or conventions also *order* pairings so that composition-of-compositions proceeds correctly. It is a foundational principle of cognitive linguistics that the pattern according to which sentence-meanings are built up to a "complete idea" has cognitive significance, even though the contrast between different "compositional paths" is not always logically apparent. For example, *her expensive diamond wedding ring* is presumably expensive *and* hers *and* made of diamonds, so the adjectives string together into a logical aggregate, but there are cognitive reasons why language (or English, at least) seems to compel a fixed order: *diamond wedding expensive her ring* is malformed, even though it has the same adjectives and noun base. Exploration of these cognitive principles is outside the scope of this chapter (*see* for instance Langacker's *Cognitive Grammar: A Basic Introduction* [29, e.g., page 320, or in general Chapters 10–11] for details), but we can observe here that most linguistic constructs demand a fixed word-order, where rearranging the sequence changes the meaning.

Moreover, the order in which words are enunciated propagates to an order among word *pairs*, and therefore among the *compositions*, which word-pairs designate. In this sense, the role of grammar is not only to specify which words pair with which, but also to induce a logical ordering amongst all the word-pairs. This appears to be the motivation for Coecke *et al.* turning to conceptual space theory: the logical ordering among word-pairs is a reflection of how meanings compose and, as such, accumulate through a sentence, leading to a complete idea. The *order* in which this composition happens suggests that the emergence of a holistic meaning occurs in stages, so that we can trace a "path" through a kind of conceptual space, where our conceptualization gets more complete and concrete as we cognitively process the sentence in its entirety. Coecke *et al.* use conceptual spaces to model the "path" that captures how information accumulates, how there is an accretion of conceptual detail, which follows the composition of individual word-pairs into a whole sentence.

This chapter will make the further assumption, in the spirit of cognitive grammar, that syntagmatic principles can be examined with an emphasis on the *epistemics* of the speaker—i.e., whomever formulates a linguistic artifact—and in particular that sentences' meanings are organized focally around *verbs*. Once again, the cognitive rationales for these assumptions are outside the scope of this chapter, but we follow linguists, such as Ronald Langacker, and note that (as cognitive artifact) any verb "profiles" an event, state, or process, and correspondingly, there is for each verb a potential propositional content, or facticity, which when concretized

[4] This is probably a compelling but not self-evident thesis, because there is a possibility that our syntactic instincts allow grammar to play these roles in a manner which is rigorous in the context of cognitive processing, but difficult to formalize in more mechanical environments, such as NLP systems.

NEUSTEIN, 978-0-323-85197-8

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 9
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

produces some sort of "complete idea."[5] For instance, with the generic concept of *moving*, there are concrete events wherein something specific moves (and so a specific time and place, or two places, an origin and destination). The cognitive acts which *profile* the given event thereby encompass a specificity, which can be expressed in propositional terms. If I see a car moving, say, I see evidence of the fact that the car has moved. Verbs are, as such, intimately linked to propositions in the sense that the concretization of a verb corresponds to the concretization of propositions; note that we could not say the same thing about other parts of speech, such as nouns. The phrase *my neighbor's black dogs*, for instance, concretizes the idea *dogs*, but does not yield a complete/specified proposition.

Earlier we intimated that the progression of sentences toward complete meanings is a matter of "accumulating information," or "accretion of detail"; from this cognitive-linguistic perspective, we can more precisely suggest that such accretion of detail is governed in particular by the tendency of sentences to converge on concrete propositions, and moreover for this convergent process to be guided specifically by the concretization of verbs. In short, the "paths" of sentences through a "space" of increasing information-content can be modeled more rigorously as progressions toward *verb* details: the accretion of information-content is first and foremost a matter of filling in details associated with verbs. At a minimum we pair verbs with a subject, and often a direct object (and sometimes also an indirect objects); we can then add further details, sometimes cued via declensions or related case-markings, specifying data such as *when*, *where*, *why*, *for whom*, *toward where*, and so forth; the verb's event (or process or state) happened or is happening. In short, "paths" in meaning space (however this should be modeled) lead to verbs; word-pairs which do not involve verbs, such as article-noun, are intermediary segments of the path, and ultimately derive their meaning from how the relevant noun connects to a verb, as subject or object (or some more peripheral case-detail, such as location).

This general picture captures (at least as a theoretical hypothesis) the overall cognitive dynamics of language-understanding, where we can visualize the cognitive processes governing sentence-interpretation as an accretion of detail organized around verbs, and the different sorts of relations verbs bear to nouns, which in turn "slot in" to expected roles vis-à-vis the corresponding verb. A ditransitive verb, for instance, presents the "expectation" of a subject-noun, object-noun, and indirect object; whichever nouns play those roles therefore fit "slots" that we perceive as needing specification, once we identify the relevant verb as ditransitive. These nouns are then linked to the verb in networks defined by such expectations, and by the distinct roles played by (e.g.) subject and object vis-à-vis a verb. These network dynamics fan out from the central verb to other linguistic elements; for instance an adjective modifies a noun, forming one step in the "path" leading to a sentences' "complete idea"; but if the sentence is well-formed this adjective-noun path will ultimately connect to a verb through a *slot* such as subject or direct- (or indirect-) object.

To give a sense of these dynamics with a concrete example, recall Chapter 6's Fig. 6.1 showing a hypothetical parse-graph of a hypothetical sentence elaborating on "went to the store" (there is no special value attached to the precise layout and terms of this graph in the current context, so we won't discuss it in detail). This diagram may convey in pictorial form the general idea of a "verb-centric" grammar and the concomitant semantics, wherein meanings are grounded in the information content they supply to a verb. Tracing paths in graphs, such as notated in (Chapter 6) Fig. 6.1, shows how word-pairs (graph edges) lead toward verb-nodes, and

[5] Again, Chapter 11 of *Cognitive Grammar: A Basic Introduction* is a good reference for this branch of Langacker's analysis; or *see* [30], [31], [22], [2], [51], [28], etc.

NEUSTEIN, 978-0-323-85197-8

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 10
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

edges in general are labeled with role-indicators (either subject/object or a declension-case) that summarize the kind of detail supplied by the relevant noun (or noun-phrase) to the target verb.

Such a perspective on natural language makes specific claims about the structural principles which need to be modeled by representations of linguistic form, e.g., descriptions of the parse-trees or parse-graphs of sentences. In particular, we have the proposal that the basic building blocks of parse-structures are word-pairs; that word-pairs can be chained together, and moreover have an overall logical order which retraces a "composition of meanings" at the semantic level; that this ordering is centered on verbs, so that the "chains" among word-pairs (and by extension the words themselves) lead to verbs; that the "links" slotted in to verb represent different roles (subject and object, most notably, but also details provided by different cases according to declension markings, where these are morphological features of the relevant natural language); and that the accretion of detail progresses to something like a determinate propositional content. Such a model can, in essence, be summarized, or formalized, via parse-graphs, where the features just outlined represent criteria on graphs, insofar as they model sentences on this paradigm. For instance, such graphs have the feature that their edges can be logically ordered, which in turn would induce orderings between the nodes spanned by an edge (given two edges incident to the same vertex, one edge is prior to the other in the ordering; as such, the non-shared vertex in the prior edge can be ordered prior to the shared vertex, whereas the non-shared vertex in the posterior edge can be ordered as posterior to the shared vertex). Moreover, following these induced edge-ordering, yields paths across the graphs, and the idea of semantics as "verb-centric" corresponds to the restriction that all such paths lead to verb-nodes.

9.2.1 The emergent syntax/semantics interface

It might seem that such a theory is still at the vague/underspecified stage without stating more explicitly what "information content" is. That could be, but such a perspective has a further dimension, which may not be immediately obvious; in particular, we are starting to analyze the dynamics that govern *why* grammar comes into effect, or emerges in its explicit form in natural language. Our framing identifies the evolutionary pressures, which appear to guide the syntax of language as conventions change over time. In other words, this is a theory not only of the explicit rules we can identify in language grammars, but also of the dynamic principles governing the manifestation of grammar as such; specifically that grammar has the role of filtering and ordering word-pairs so as to map sentences onto a kind of "space" endowed with a notion of information content, where meaning-composition corresponds (or "maps") to paths in this space that lead to "complete ideas" (e.g., propositions). Coecke *et al.* capture these mappings in category theoretic terms: "functors" map paths in a space defined by formal *grammars* onto paths in a space defined (to some approximation) by conceptual space theory, with the idea that the former "syntactic" paths somehow guide us (or mathematically model the cognitive processes which guide us) to grasp or follow the corresponding "semantic" paths in "conceptual" space.

Apart from the cognitive and/or computational merits of this theory, it also can potentially lead to new perspectives on the dynamics underlying grammar as such, as just intimated. One way to examine this is to consider the contrast between *syntax*, or the explicit (and to some degree statically analyzable) grammatic rules of a language, with (as it may be called) the *syntagmatics*, or *patterns of lateral organization* observable in language. The term *syntagmatics* is more associated with philosophical linguistics than computational NLP, but in general it

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 11
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

tends to connote an emphasis less on the explicit syntax of language than on the principles guiding the emergence of grammar: the syntagmatic "pole" of language is the order we perceive in how word-sequences follow a consequential progression, so that word-order is neither random nor (in general) freely modified; the sequencing of words conveys meaning no less than the words themselves. The principles governing this ordering are a kind of dynamic arena, where specific syntactic rules can be defined, so we address the more "syntagmatic" aspects of language if we investigate the overall dynamics of syntax, as opposed to specific syntactic rules/conventions. Or at least, in Chapter 6, we adopted this kind of usage, and refer to *syntagmatics* as the dynamic principles explaining the cognitive and structural principles guiding the emergence of syntactic rules (where *syntax* as such is less abstractly focused on concrete grammars).

Our hypothesis, to summarize, is that syntagnamic dynamics are driven by the interplay of syntax and semantics vis-à-vis information content: syntactic rules emerge under pressures to engender semantic constructions, which embody amplifications of information content in well-structured ways; there is a clear sense of how each part of a construction elevates the information-content as a whole. In the context of conceptual spaces, we would argue that the quantitative dimensions, which are internally invoked by particular concepts' cognitive "footprints," are indeed part of such information content—they help articulate how each concept add its own detail to a situational whole—but that conceptual combinations embodied by multi-part constructions should not be modeled (in the usual case) simply by juxtaposing (whether via union or intersection) dimensions internal to every concept spanned by the governing construction. Instead, qualitative dimensions combine with situational roles in potentially complex ways.[6]

This model, we would argue, represents a semantic paradigm general enough to apply both to natural language and formal languages (as well as data semantics). Formal semantics in these contexts need to consider the internal dimensions of particular "concepts" (or whatever notion plays an analogous role, such as types, or ontology classes) as well as role-inflected aggregations, where multiple concepts are integrated (be this in procedures, multi-part relations, pre-persistent object representations, type-to-visual-object mappings, and other formal metastructures; recall our "semiotic saltire" outline). While the terms of such a model are not endemic to conceptual spaces, that theory does serve as a rich starting-point for intuitively driving such a role-oriented picture, because conceptual spaces provide a good case study in how roles as well as intra-conceptual dimensions and details (which present different concept-blending options in constructional contexts) determine "paths" engendered by syntagmatic constructions.

The remainder of this chapter will elaborate on the framework just referenced through various hypergraph-oriented representations of code and/or data structures, including multi-relations, grounded serialization, and type-persistence models.

9.3 Conceptual and thematic roles

Chapter 6's discussion of *multi-relations* (relations with multiple parts and components),

[6] We could also say *thematic relations* as one way to characterize situational roles, adopting terms from Case Grammar and related fields, although outside that specific context the generic term *thematic relation* can take on many information relations, so it may be suboptimal for analyses outside of those targeted directly at case grammar, inflectional syntax, and so forth.

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 12
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

together with distinct *roles* attributed to participants in the relation, alluded to the processes whereby multi-relations contribute multiple facets of information to the knowledge-contexts, where they are believed/asserted. Consider an assertion that (continuing that chapter's example) *John* divorced *Jane* (together with details such as divorce and marriage dates). Clearly that relation, to the degree that it is taken and used as fact in some relevant reasoning context, provides specific information. The *nature* of this information depends on our reasoning purposes, which (in a graph-modeling context) amounts to how we are traversing a graph.

If, say, we are "visiting" the John node and wish to learn his ex-wife's name, the "divorce" multi-relation supports that query (by stepping through the relation to the *Jane* node). The multi-relation (at least with data mentioned in Chapter 6) also supports the step of learning a divorce *date* by traversing to *that* (date) node. Similarly, it supports the confirmation that John *has been* married, because there is a path from John to a *marriage* node (via the *divorce* node).

In short, though we speak generically of "information content," we can be more precise about the operations *entailed* by information content by examining the processes afforded by information-encodings (such as via knowledge-graphs). The information content contained within a graph element (e.g., a "divorce" node) corresponds to the traversal-options accorded to paths through or across that element, and how these paths yield data at the sites which they visit. Implicit information content is "accessed"—made *explicit*—by following paths in an information-space/knowledge-space (any space encoding an aggregate body of information/knowledge). Such a notion of *paths* falls quite naturally out of knowledge engineering, and we can consider how it resonates with the analogous picture of *semantic paths* in the context of natural language.

In Chapter 6, we also suggested that multi-relations are in a sense "dual" to *procedures*. Steps in syntagmatic graphs involve accretion of detail targeted at supplying enough information to run a procedure. As a precondition for calling a procedure, one must bind values to each required parameter, so all of these parameters point toward the procedure in the sense of supplying data prerequisite for the culminating procedure-call step to proceed. Collectively, each input parameter is then a kind of aggregate step, or a collation of steps into a multi-part step, where each step's purpose must be satisfied. Syntagmatic graphs can encode semantic requirements if we argue that procedure-node neighborhoods encode the totality of information, which must be available prior to a procedure-call, and that a collection of individual steps, each providing *one* point of info (one parameter-binding), logically entail a kind of *multi*-step, which is the epistemic phenomenon of achieving the full information content requisite for a given continuation (e.g., a procedure-call).

The logic of this picture is structurally dual in several different respects for multi-relations in lieu of procedures. A typical multi-relation is traversed from one initial node (one *input*, from the perspective of a specific traversal occasion) and can branch into multiple destinations (multiple outputs); this is dual to procedures taking (in general) multiple inputs to one output. These outputs are possible steps that *may* be followed (so we can picture the steps unified by a disjunction, a modal combination of steps being the set of steps which *may* be taken), whereas procedure inputs *must* be supplied (a step-*conjunction*). In general, a procedure's output for a given input is not known *a priori* (this is why the procedure is called in the first place), so this response is located *in the future* relative to the procedure call; whereas the information contained in a multi-relation has been provided *in the past* (stored in the database/knowledge-base), but now has to be retrieved again. In short, introducing modal and temporal operators serves to sharpen the duality of procedures and multi-relations.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

Since duality implies structural similarity, these considerations suggest that syntagmatic graphs are useful for modeling multi-relations as well as procedures, an idea that is borne out by considering how multi-relations aggregate information content. We can adopt the convention that nodes *point in* to a multi-relation node, since each node in the relation's neighborhood supplies data: in the *divorce* example, *John*, *Jane*, *2015*, and *2020* all contribute a field to the data structure embodying their divorce. So we can traverse from the suppliers of data to the target, where that data is synthesized. On the other hand, if we are *traversing* the graph (some time later than when the data is first assembled), we move from the multi-relation node to one of its neighbors (from *John*, say, to *divorce*, and then *Jane*, to learn that Jane was whom John divorced). In short, directions associated with *traversal* are often inversions of those associated with *information content*. This is analogous to the input/output parameter distinction for procedure-neighborhoods. As we argued in Chapter 6, even output nodes *contribute information to* a procedure (e.g., a type-constraint on its return type), so as a *static* model (e.g., for a compiler), a return-value node provides information *to* the procedure-node. However, in the *dynamic* case of actually stepping through a source-code graph, we can step *out* from the procedure to its output node.

In general, the *effects* of a procedure propagate in inverse directions from the accretion of detail, which preconditions the procedure. All parameters affected (both input and output) supply information (type-constraints if not actual values) to the procedure, while at the same time the procedure may affect (by overwriting) those parameters. In fact, in type-and-effect-systems, one commonly refers to *reading* a value (not just *writing*) as an *effect*, because, rigorously speaking, reading values change the state of their carriers (for example, because the area of contexts where the value is known is then enlarged, which is consequential in logical systems for cybersecurity, say). Thus anything which supplies information content to a procedure may potentially be affected *by* the procedure, with the definition of "affected by," depending on the "effect system" one chooses for modeling procedural side-effects. However, following such effects is not a *traversal step* along graph-paths; rather, it involves tracing evolution in the *program state* which is modeled by source-code graphs. Effect-directions are therefore not explicit structures in the graph, but indirect pathways introduced via code-graphs' semantic interpretation.

In the case of multi-relations, a given relation-node centers a neighborhood of multiple other nodes, each supplying some information into an aggregate. The static picture encoded in such a neighborhood involves how those disparate data points are merged into a data structure embodied by the central node (this expresses the aspect of multi-relations, where they resemble *objects*). On the other hand, dynamically, a multi-node is traversed *from* a neighboring node *to* a neighboring node (this expresses the aspect where multi-relations act more like "edge tangles"), so "dynamic" traversal represents different path-directions than "static" traversal, similarly to procedures.

The question of *which* node will step "into" the multi-relation, and which will be the destination, depends on dynamic context. For procedures, the totality of information available in the system once calculations are performed is greater than the information statically modeled by code graphs. The graphs capture open-ended information possibilities, but closed traversal options, in that one can only "step into" a procedure (using terms canonical to debuggers) after multiple prior steps (for parameter-bindings) are completed, and can only "step out of" the procedure to a specific output node. Dually, for multi-relations, the information is not open-ended (it is fixed ahead of time), but the dynamics of how the multi-relation node is traversed are open-ended in that they are context-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

dependent on the state of whatever software if "visiting" the relevant data/information space.

Both procedures and multi-relations reveal structural parallels in terms of static *a priori* data contrasts with dynamic context-specific data, and in terms of how these contrasts align with path-directions on the graphs. In both forms of representation, there is a distinction between explicit graph-structure (which supports certain kinds of paths) and dynamic context, which provides a semantic interpretation to the graph yielding alternative paths (e.g., following procedures to their side-effects). Such indirect paths are not fully explicated in the procedural context, until there is a dynamic context (an actual procedure call), where, for instance, side-effects can be observed. Analogously, paths across a multi-relation to neighboring nodes (the edge-tangle guise of the relation) may only be concretely available with a running context, wherein one is visiting a specific input node to the relation. In the case of multi-relations, *roles* can provide guides for these dynamic contexts, consolidating which path *out* of the multi-relation is appropriate for the context.

In effect, it is only through roles that multi-relations can be a basis for well-defined traversal in the first place, because roles specify how to *leave* the relation-node once it is *entered*. For example, the "divorce date" role permits a visitor to step from John to *divorce*, and then 2020, to learn the desired datum; that 2020 was when John divorced. Roles provide sufficient structure for multi-relations to be nexus-points from the perspective of graph-traversal.

We will argue that roles can potentially be used in a *procedural* context for, in effect, a dual purpose, in the sense that *accumulating information content* (adding in data) is dual to *accessing* information content ("pulling out" data). Specifically, roles can play a disambiguating service with respect to how information content *plugs in* to a procedure by analogy to how roles "disambiguate" the range of traversal path-options for multi-relations when their information is accessed. This effect is more pronounced in natural language than programming languages, we will argue, but a similar notion of "roles" can potentially be useful in a programming context.

9.3.1 Disjoint conceptual spaces

Above we suggested that conceptual space theory has perhaps overemphasized the "narrowing" logic of concept-juxtapositions, and implied that we can have rigorous accounts of information content that do not depend on scope-narrowing to realize the paradigm wherein semantic paths converge to precise meanings. Although scope-narrowing is one manifestation of a dynamics tailored to greater precision, what is really at stake in semantic paths, according to the syntax-semantics interface as we outlined it earlier, is that semantic paths lead toward greater information content. In other words, the juxtaposition of two concepts should produce greater information content than either concept alone; however, this model does not require that such amplification occur, *because* the concepts' narrow each other's scope.

In this chapter we will turn to *roles* in the AI sense, and indirectly to analogous ideas in conceptual role semantics, to provide an alternative model of information accretion. Two concepts can be juxtaposed to an effect of *more information*, insofar as they have *distinct roles*, even if the concepts are not mutually narrowing. In *I drove her to the store*, the respective concepts (*I*, *drove*, *her*, *store*) do not appear to combine by limiting each other's scope, but rather by offering different roles to the overall meaning (agentive, patientive, locative). To the degree that we can perceive "spaces" around these respective contexts, it seems that the holistic meaning emerges not from a blending of the respective spaces, but from collating them *in an ordered fashion* by indexing in terms of roles, so to speak.

Conceptual space theory appears motivated by the notion that concept-juxtapositions acquire specificity by how conceptual spaces in-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

teract. Word-jumbles, which would seem to tokenize random assortments of disparate concepts (and their spaces) do not carry substantive meaning apparently, because they do not permit the concepts to combine in productive ways. This would obviously be the case with pairings having no interpretive path linking one concept to another (*curried theorems*, say). But we can also imagine encountering words in a context, where one could suppress the idea that they have any syntactic linkage (consider a scattering of newspaper clippings on the floor). So (say) *red* and *square* taken as just happening to mention two different concepts does not contain any specificity, but the construction *red square* implies that the two concepts are to be *combined*, and this combination is interpreted by intersections and unions of dimensions.

When concepts are combined by attributing them different roles, the pairing is neither completely free-form (they are still a linguistic *construction*, not a random jumble of words), but nor is it dependent on dimension intersection/union for significance. For example, *drove her* is neither an unstructured tokening of two words without any implication that they form a linguistic unit, nor the construction of a fused concept, such as *red square*. In effect, roles (subject/object/location/benefactive, and so forth) provide a compositional principle *alternative to* the quantitative models typically advanced by conceptual space theory.

Coecke *et al.*'s strategy, wherein conceptual spaces provide a semantic paradigm above pregroup grammars, for example, applies most directly to simple verb-constructions with subjects and objects. In such contexts, at least for "**SVO**" languages with the verb (for normal clauses)[7] positioned between its subject and object, pregroup structures formally predict the type-reduction of the verb to a proposition, given left-combination and right-combinations with nouns. As with combinatory categorial grammar (*see* [5], [36], or [43]), reductions based on left and right adjacency can thereby formalize conceptual constraints on our semantic interpretations of the relevant clause: reduction of the verb-construction to a proposition captures how our attributing parts of speech to component words guides understanding of how the components should fit together; in the case of a finite clause, they should combine into a propositionally complete idea. Type reductions at the syntactic level thereby map onto communicated meanings at the semantic level, or at least onto interpretive parameters, which allow those meanings to be deciphered.

This picture is more complex, however, when we consider the full range of verbs' "theta roles" (which extend beyond subject/object to include patient and agent roles and/or indirect objects), and then to the full range of thematic roles available through nested clauses and/or case-marking. With as many as three theta roles and potentially several other thematic roles, verbs do not just "look" left and right for their **S** and **O**; they also accept links to other words or phrases, which are neither subject nor object, but instead provide some other kind of detail. All of this complicates the pregroup-grammar model of type reductions (*see* [27, especially Chapter 4] for a formal statement of similar issues).

Additional complications arise from the possibility of "movement" in theta roles, canonically (in English) the transpositions exemplified by sentences such as *I gave the book to John* and *I gave John the book*. Also, *polyvalent* verbs present a problem, because we need to understand which "theta frame" (determining the roles related to agent/patient status we expect to be linked to the verb) applies to each such verb in a given usage; and thematic roles can be "optional" in that we hear some thematic clauses as required by the verb-sense in force, whereas others are supplemental details. The normal sense of *put*, for instance, requires a patient *and* a location (*I put her book on the table*), whereas for, say, *read*, the

[7] Not questions, and so forth.

NEUSTEIN, 978-0-323-85197-8

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 16
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

location is supplemental (*I read her book on the train*).

All of these phrase-structure options force verb-clause theories to have more structure than left- and right-adjoints (as in pregroup grammars) alone, which complicate the calculus of type-reductions over clauses. These may not be intractable problems for approaches such as those of Coecke *et al.* from the *syntactic* side, since multiple thematic roles can always be treated as additional verb-arguments; instead of "left" and "right," we can posit a trivalent or multi-valent set of "directions," which link the verb to different subsidiaries, indexed by thematic role rather than by directions in the surface expression (as an aside, needing to accommodate these multiple "directions" is further motivation for our "syntagmatic" graph paradigm, where all edges point *toward* procedure nodes).[8] Type reduction is then a multi-stage process that depends on contributions from each component thematically related to the verb.

The question is how to accommodate this more complex notion of type-reduction to Coecke *et al.*'s derivation of semantic constructions being *engendered* by syntactic constructions, with the meaning-combination being effectively an "image" of a mathematically determined map *from* the syntactic construction. The authors show us how this syntax/semantics interface works in the context of simple bivalent verbs, where the relevant semantic "image" is a blend of two concepts. In the case of multivalent verbs, we have multiple concepts, which have to be merged together according to thematic roles. And this is precisely the kind of scenario where we have argued that concepts tend to co-exist in situational models with their own internal details, rather than blend into a dimensionally integrated whole. In *I drove the van to Philly*, say, the *van* and *Philly* are situationally autonomous in the sense that one could plausibly drive the former many places, and plausibly travel to Philly via many means. The spaces "around" the patient and location roles are largely disjoint.[9]

In this sort of example, any quantitative dimensions endemic to the respective concepts are in most cases autonomous, as if the sortal effects of their respective roles *inhibits* dimensional convolution. At most, we could say that the dimensions are "latent" and provide cognitive background that can *potentially* interact. An example would be *I took the express train to Philly in under two hours*. Here the geospatial dimension ambient to conceptualizing Philadelphia intersects with the figurative "dimension," involving different kinds of trains, because the sentence intimates that the length of the trip was affected by the train being of the "express" variety. But opportunities for dimensional interactions along these lines are latent, rather than intrinsic to role-indexed constructions; in the usual case different concepts (and their dimensions) remain mutually autonomous. To the degree that this is true, quantitative models of concept blending need to be supplanted with situational models that place greater emphasis on how multiple role-construals synthesize into cognitive schemata. It is not clear that this happens on a linguistic level at all, rather than a prelinguistic

[8] Because multivalent verbs imply multiple directions of "adjacency," rather than the binary options of "input" and "output." Edge directions therefore no longer line up neatly with "adjacency" directions; the latter is an additional structural detail *on* the graph-representation, so edge-direction is not interpreted as fixing the imputed direction in this sense. Instead, edge-direction is an artifact of the graph on the basic syntagmatic level, driven by the convention that procedure-nodes are canonically targets, but not sources, which in turn is motivated by the goal of partitioning graphs into neighborhoods with few edges crossing between them.

[9] For the sake of discussion, we assume the role of *van* is *patientive* when it is a direct object (what was driven), though perhaps vis-à-vis cognitive framing, its signified role is more instrumentive, *except for* cases where the specific purpose of the drive was to deposit the van; otherwise the vehicle is merely the means selected for travel. Syntactically, though, the van is only presented in an instrumentive mode when it is not a direct object (*I drove them to Philly in the van*).

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 17
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

level of cognitive construals and anticipations of the unfolding affairs around us.

Furthermore, insofar as one goal of conceptual spaces is to model similarity and prototype effects, it seems hard to isolate thematic roles from accounts of how we judge the scope and applicability of concepts. In an analysis utilizing *birds* as a case study, for instance, [38, page 158] constructs a "bird space," whose regions correspond to types of birds, analogous to color-concepts being regions in color space. Moreover, there are exemplars within this space such that *robins*, for example, are more prototypical than *penguins* or *ostriches*.

We can, of course, compare birds in different ways (size, color, quickness, etc.), and it is likely that we do indeed as a community share an exemplary bird-notion, which some bird species embody more than others. These two factors make pictures of a Euclidean "bird space" intuitively appealing, because they capture all together the geometric possibility of a prototype, the gradations of similarity/dissimilarity and prototypicality, and the multi-faceted nature of "inter-bird" comparisons. But the success of such a picture in iconifying these intuitions is not *prima facie* evidence that most intra-concept comparisons can be situated along a numeric axis (like size), or even domains of multiple axis-dimensions (like color). Penguins and ostriches are atypical birds, largely because they do not fly, and *flight/non-flight* is not a quantifiable domain in any straightforward sense; instead, this distinction depends on our appraising birds' traits and behaviors as functionality organized adaptations to their environment.

Functional criteria along these lines tend to be captured conceptually via thematic roles more than via grade scales. Cars, trains, and buses are all plausible vehicles to take us somewhere, but they are not geometrically comparable, such as red, green, and blue; rather, they provide somewhat different enactive affordances, and each subconcept presents, via its tokens, a distinct genre of functionally organized system. Our background knowledge of how maneuvering into and within a *car* differs from *bus* and *train* is most likely to be exercised in the course of sentences, where these concepts would be pressed into service as instrumentives for verbs of geospatial motion. Our cognitive resources for construing such vehicles as "movement tools" supply the scaffolding wherein instrumental constructions along these lines are given sense: what it means for a car/bus/train to be an "instrument." Apparently the same prelinguistic knowledge is manifest in our grasping the similarities and differences between car/bus/train as "vehicle" subtypes. So it seems hard to model the conceptual space wherein the subtypes acquire their patterns of similarity/dissimilarity (and their respective prototypic cores) without focusing on their distinct patterns of functional integration (in terms of how they operate and in terms of how we make use of them).

Though we can abstractly imagine a "space of possible functional organizations" wherein cars, buses, and trains would have their own regions, in reality such a space would seem to be at most an emergent summary of our background knowledge vis-à-vis modes of transport: the basis for our contrasting car/bus/train is all the accumulated knowledge we have of their basic workings, the mental "scripts" we subconsciously follow when entering/boarding a vehicle, or interpreting its movement or planning a trip, and so forth; we do not rely on immediate perceptual cues, such as size or color to establish inter-token comparisons, in general. Instead, such comparisons derive from situational reasoning and anticipation as we are either in the process of traveling via car/bus/train or planning to do so.

We do not intend to rigidly pursue this point, since one can find implicit counter-arguments in the conceptual space literature: Gärdenfors and others, for instance, have specifically written about modeling phenomena, such as functional organization and such as spatial paths

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

via (quantitative) conceptual spaces [26] (*see also* [11], [32], [15], [55], [8, especially section 3], or [25, especially pages 5–7]).[10] So there is precedent for modeling (say) verb-plus-locative constructions as indeed quantifiable narrowings (akin to *dark red*), rather than as the basically unblended juxtaposition of two distinct conceptual spaces [56]. Accepting one or another analytic strategy presumably depends on how strongly one is convinced by a Gärdenfors-style encoding of themes such as spatial paths, movements, situational contexts, prototype/borderline contrasts, etc., as quantifiable structures (with sufficient abstractness, almost anything can be seen as part of an at least metaphorically quantitative "space," but the degree of abstraction involved may seem to weaken the force of the overarching model, even though we have to grant theories the option of establishing claims in more concrete cases, where the theoretical commitments may be more clearly demonstrated, such as color-spaces as canonical cases of perceptual dimensions, and then generalize to analyses, where the terms of the theory need to be applied more abstractly or obliquely). Nevertheless, we can accept the premise that certain cognitive details pertaining to individual concepts have (to varying degrees) quantitative form without arguing that concept *combination*, and therefore meaning-composition (in linguistic contexts) is often defined by a quantitative merger of the two concepts involved.

In *drove to the store* one can, for example, certainly argue that *to the store* has a scope-narrowing effect on *drove*; by itself the verb is open-ended, but the subsequent phrase narrows the verb's scope and concretizes its profiling by supplying a destination for *drove*. This narrowing, however, appears to be a matter of information content being supplied to fill in an abstract slot with concrete details (the "where?"/"to where?" of *drive*); i.e., a narrowing from abstract (compatible with many possible states of affairs) to concrete (empirically specific). This sense of narrowing is different from the refinement effects of, say, *dark red*, and it is hard to see how *to the store* could *quantitatively* alter *drove* akin to how *dark* alters *red*.

In reading Gärdenfors-inspired literature on event semantics one might get the impression that in different places thematic roles (i.e., the concepts that play them) are sometimes understood to *have* quantitative dimensions (e.g., a locative is the culmination of a spatial movement, and so picks up the dimensions of the space wherein such movement occurs) and sometimes understood to *be* dimensions; perhaps reflecting how proponents of this theory have not converged on a paradigm for incorporating event/situational semantics. In, say, *Last week I took some students to Philly by train to attend a conference*, the multiple thematic roles provide a kind of mental "checklist" of details the speaker feels relevant enough to warrant men-

[10] One problem examined by the authors just cited is that an event construed in different ways could be taken instead—amongst linguists anyhow trying to decide how best to set up philosophical parameters on event semantics—as more than one event, or not. Does *drive to Philly* name the same event as *drive to the conference* in cases where someone does the former so as to do the latter? There are counterfactuals that might suggest that these are in fact two different albeit strongly overlapping events: e.g., were they stuck in traffic and late to the conference; one could still say they were *in Philly* on time. On the other hand, counterfactuals are, indeed, contrary to fact: arguably *one* event *could potentially* have been two events, were some factor to intervene. Event-semantics literature does not seem to settle such issues. But by extension the questions involve seem to impinge on whether spacetime and force-dynamic construals are *sufficient* to characterize events or whether we *also* need some notion of thematic roles, which would help negotiate counterfactuals and other complicating factors. E.g., if we read *Philly* as meaning the *patient* of the drive-to-conference action, then they have not actually driven "to Philly" if they are within that city's municipal boundaries, but not at their destination, since in the context of that particular sentence Philly plays a thematic role, which is not satisfied by just any trajectory, which happens to terminate somewhere in the city. It is not clear how that treatment of the counterfactual would be possible *without* thematic roles, or what a more "quantitative" gloss on the counterfactual would look like.

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 19
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

tion. We can picture the full set of (sufficiently relevant) details as a kind of virtual dimension characterizing the verb (no less than the quantitative dimensions that would literally model the movement, i.e., a spatial trajectory ending in Philly). According to Gärdenfors, indeed, "The cognitive structure of events is relational, gluing together objects, actions and locations" [25, page 6], which sounds as if the theory pictures events as bundling components, each having distinct cognitive status. The *dimensions* of such a construction would seem to be the components themselves, that provide relata "glued together" into an aggregate. Though not explicitly conceptual space-oriented, but in a theory with detailed analyses of dimensional structures, Lucas Champollion invites a similar reading in summarizing how (in his "strata" theory) "events ... are thought of as occupying regions in an abstract space whose dimensions specify, among others, their spatial and temporal extent" where "thematic roles and measure functions [are also] among the dimensions of this abstract space." [12, page 127] In other words, situations entail a "space of thematic roles" (including but not limited to *theta* roles) and players of individual roles (locative, say) are akin to "points" in this space.

But elsewhere Gärdenfors uses language such as "the force exerted by the agent will be modified by the instrument and thus different from the force vector affecting the patient" [56, page 21], which sounds as if subject and instrument are not *related* so much as "fused" into a "single" force-vector which is the one "affecting the patient." If *that* is the theory's archetype, then thematic roles would paradigmatically blend into quantifiable domains, so the "points" in the blended conceptual space would not be enumerative discretized spaces of possible thematic roles, but would be more mathematical and quantitative, e.g., force-vectors, which seems a different picture than "objects, actions and locations" "glued together" (at least insofar as the "glue" is a situational inter-coherence innate to thematic roles).

Our point is not that Gärdenfors is being inconsistent, but rather that neither thematic relations nor quantitative dimensions alone can model event-semantics: roles and dimensions situationally interoperate in many different ways. For example, Champollion's strata theory, which we just mentioned, is a case study in how formal semantics can integrate the qualitative aspects of thematic roles with the quantitative features of conceptual dimensions characterizing the concept-instances, which play those toles.[11]

In general, we would argue that roles more often than not *inhibit dimensional interactions*, so that quantitative representations, such as Gärdenfors's "two-vector" system typically apply

[11] Dimensional structures help to organize semantic constructions in a number of patterns (such as partitives, mereologically inflected reference, and quantifier-scope): attributes such as mereology and granularity (as well as quantitative notions of magnitude and comparison operators) can all play a role in the semantic rules governing part/whole relations (in contexts such as *each of the ...* or *a few of the ...*), gradations (*more ... than ...*), effects related to the completion or partiality of events (*for an hour* vs. *in an hour*), and so forth. All of this builds up a theory of dimensional structures as constituents of referential scenarios, whose linguistic encodings appear to be conventionalized within construction templates involving partitives, quantifiers, comparatives, etc. However, Champollion explicitly allows for these concerns to be applicable to different thematic roles within an overall event semantics: to the degree that (glossing his arguments with terms other that his more precise, but technical alternatives) *dimensional structures* offer a kind of *semantic frame* governing recurring quantity/mereology-related construction-patterns, then such a frame "distributes over" (which *is* his term) thematic roles. Intuitively, each player of a role (each "theta," in his terms) can have its own such semantic frame. Dimensional structures are, that is, "θ-indexed" in the sense that θ becomes one "parameter" (selecting the relevant thematic role) alongside parameters defining dimension-related structures such as granularity (*see* [10, page 33]). This would be an example of how dimensional structures come into play *in the scope of* one specific thematic role, so that the totality of dimensions appertaining to a semantic construction—and to the semantics of an event—can be defined only by considering *both* the qualitative distinctions between roles *and* the quantitative aspects of concepts through which we make sense of the entities playing those roles.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

only to agent and patient roles; other thematic roles should be analyzed situationally, rather than as "vectors" (although they may have latent dimensions, which can *potentially* be semantically convoluted with the central agent/patient schema in some cases, as in our express-train example, which should be taken as part of the semantic arsenal comprising a clause's overall meaning).

For the most overarching analysis (such details as "latent" comparatives aside), rather than treating concept-combinations as meaning-effects, which have to be modeled numerically, we can instead interpret concept-constructions as organized situationally: different concepts lend different sorts of detail, each of which add concreteness and specificity, leading from situational abstractness or prototypes to concrete states of affairs. Ordering concepts by their conceptual *roles* can thereby take the place of quantitative modeling for conceptual blends. This role-inflectional approach—which we might characterize in terms of *qualitative*, rather than *quantitative* concept-combination—is arguably better motivated by natural language. The fundamental aspect of qualitative combination in this sense is that the spaces of two concepts typically do *not* blend, but rather remain intellectually isolated, with the rationale of the concept combination being articulated through distinct syntactic and (correlatively) situational roles, rather than through dimensional interconnections. Dimensional structures for individual concepts may still be important as part of our holistic semantic assessments, but in many contexts the situational composition fusing multiple concepts is more significant for overall meaning than the precision afforded by concepts' quantitative dimensions (e.g., the color-space boundaries marked by the concept *red*).

Autonomous concept-blends still, we would argue, evince certain (often partly metrizable) dimensions; indeed, concepts "glued together" have a larger space of domains and dimensions than concepts individually. Given a kernel event, such as *travel to Philly*, we can elaborate on many dimensions detailing that situation, such as *how long*, or *where exactly*, or *when*; some of these details take a scalar or geometric form. But here we transition from dimensional structures *constituting* constructional meanings (as is arguably the case for simple aggregates, such as *red square*) to dimensions being *available for elaboration*, which implies that they supplement a completed signifying process, which the dimensions themselves (or any interactions and convolutions thereof) do not fully explicate.

Indeed, to the degree that concepts within a construction are autonomous, it is not immediately evident what there is to analyze; we can always say that language-users "glue" concepts together in the mind, and that they are cued about concepts' respective roles by morphosyntax (which so to speak "sieves" concepts into distinct thematic roles), but that sounds like simply restating the explanandum, rather than explaining it. The charge for the *syntactic* side of the theory is clearer, because linguists have to demonstrate how clauses are parsed into role-attributions in the first place. We claim that a thorough account of such a process can be derived via expanding from pregroup grammars to hypergraphs, where graph edges track (collections of) nodes to corresponding verbs via the roles they carry.

One can indeed find apparent rules or conventions in clause-structure, which appear to govern this process, such as patterns in how thematic roles become registered via theta-roles or relegated to secondary status, and the relation of roles to speaker epistemics. Such observations can found a bonafide "theory" of clausal construction on the *syntactic* side. But semantically, the more that concepts are distinguished in a construction by playing distinct roles, the harder it is to find convincing formulae reducing their aggregation to some formalizable meaning-generator, as opposed to invoking a largely prelinguistic "situational cognition" outside the scope of (linguistic) analysis.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

We contend that approaching semantic constructions from the perspective of information-delta *paths* sheds some light on this situation, and can form the core of a theory, which does more than hand-wave at the problems of role-indexed semantic constructions. Even within such a theory, nonetheless, we should be receptive to the possibility that unpacking semantic constructions really *is* in many ways pre-linguistic; if nothing else, a theory can demarcate the boundary between processes endemic to language-understanding, as they play out in the context of semantic constructions, and those which defer to "situational cognition," something perhaps too subtle for linguistic analysis, and, arguably, too "human" for "artificial" intelligence.

Yet what we *can* do, even in the confines of linguistics proper, is identify where patterns in situational reasoning (including intersubjective "theory of other minds" effects) appear to drive linguistic (and pragmatic/discursive) conventions, for example in how clause-constructions are organized around the "speaker's point of view," and the need to establish common reference points and situational framings between all participants in a linguistic context. That still leaves a lot for linguistics to talk about. Even if some majority of cognitive situational process is pre-linguistic, there are still many different mental operations which could be pressed into service to construe the affairs around us. Language surely uses semantic and syntactic cues to activate certain prelinguistic capabilities from the full set that could be potentially triggered within ambient situations.[12]

A perspective which emphasizes concepts' situational roles can thus leverage many facets of conceptual space theory, even while de-emphasizing mathematical accounts of concept-blending, and moreover would still be consistent with data-modeling and meta-scientific applications of conceptual spaces as found in projects such as Conceptual space markup language. That is, a *conceptual role* based framework can be one way to integrate this data-modeling branch of conceptual space theory with the formal-linguistic considerations prioritized by (e.g.) Coecke *et al.*.[13] It is in this guise, one can argue, that conceptual space theory as a model for *natural language* semantics can also be a reference-point for settings such as data and code semantics, which lack the former's situational nuance.

There is, however, one useful analogy which may be drawn between the natural-language and data/code semantics cases. Consider the basic idea that role-indexed constructions tend by default to allow component concepts' relative autonomy (by contrast to conceptual blends without obvious thematic-role decompositions, such as *red square* or *dark red*). Due to such autonomy the ultimate "meaning" of concepts' combinations might not be constructed through linguistic means at all, but rather defer to cognitive-situational faculties. The relevant analogy in data semantics is that the full semantic constitution of a data space—of the types and schemas which are embodied by, instantiated in, or constraints on a data set or database—cannot be defined through schematic declarations or axioms alone (e.g., web ontologies). Data sets' full semantic import derives from how the software components that use them replicate or analyze some real-world situations, which ultimately depends on *procedures implemented* more than

[12] Recurring patterns in situational construals, perhaps especially *perceptual* construals, can then be seen to engender specific types of "semantic frames" (for force dynamics, mereology-related as mentioned above vis-à-vis Champollion, modality (in the sense of modal logic) and counterfactuals, belief-attributions and "other minds," etc.) which encapsulate how language via construction-templates and patterns can invoke cognitive faculties, not just isolated mental processes but interconnected sets of logical and operational schema through which we organize conscious experience.

[13] *See*, e.g., [39], [35], [45], or [42] for overviews of "Conceptual Role Semantics" in particular.

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:56 • p. 22
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

on static representations, such as data fields or inter-object relations. Thus data modeling is only preliminary to code implementation, and the *semantics* of data models lies largely in how they can guide, and then make reference to (as data-structural specifications) implementation features and procedures. We would argue that this is roughly analogous to how features of linguistic constructions both *refer* and *defer* to "cognitive procedures," which provide our ambient situational reasoning prerequisite for language.

9.3.2 Conceptual spaces and scientific data

When multiple concepts are merged into meaningful constructions, each concept contributes distinct information, which collectively produce a larger information content. Without referring to information content specifically, conceptual space theory alludes to this process and presents one version of how it operates, particularly in the context of conceptual blends: blended concepts (such as *dark red* or *red square*) quantitatively mix their concept-components to procure a more narrowly extended (and thus more information-bearing) prototype than each concept on its own. As we have argued, however, we can continue the general idea of information-content amplification without restricting analyses to these blend-cases; for example, concepts combined qualitatively assemble information content by organizing concepts' contributions in terms of situational roles more than interpenetrations of dimensional structures. Yet however we figure concept-combination, the crucial detail is marking how such combinations use the coordination of multiple concepts to augment levels of information content against that of concepts individually. This foundational question appears to lie at the core of conceptual space theory as a linguistic semantics.

Not long after Gärdenfors's initial language-focused publications, there emerged some follow-up research, which shifted emphasis from the semantics of natural language to that of scientific theories and scientific data. This included formulation of conceptual space markup language (CSML) as a data-description framework, with an emphasis on conceptually and statistically rigorous documentation of the parameters, which collectively define a scientific theory or model. The CSML language is a rigorous and well-motivated approach, which deserves to be widely adopted in some fashion, particularly given the more recent emphasis on data sharing and research transparency, which is even more pronounced now than when CSML was first published, as well as the presence of more recent data-sharing protocols that could be integrated with CSML, such as the Digital Curation Center (DCC) lifecycle [48], [50], SciXML [16], [44], [46], [33], [53], [54], IeXML [37], [40], [47], the suite of MIBBI (minimum information for biological and biomedical investigations) guidelines [52], [34], and Stuart Chalk's **SciData** Ontologies [9].

With that said, although formalizations of conceptual spaces such as CSML fit in well with these various attempts to standardize scientific data-sharing, much of CSML's details or vocabulary is not endemic to conceptual space theory *per se*. For example, conceptual space theory places particular emphasis on concepts' (and by extension scientific models') domain and dimensional properties, which in the data-modeling context inspires especially rigorous attention to phenomena such as the statistical scale (nominal, ordinal, interval, ratio), measurement units, ranges, and autonomy/correlation among different modeling parameters. However, documenting such statistical qualities of a data set is hardly specific to conceptual space theory.

On the other hand, notions such as "contrast classes" and the partitioning of a larger conceptual space into regions associated with concept-prototypes *are* more specifically rooted in Gärdenfors's original theory, but such techniques seem more applicable for use-cases, such as NLP-driven text mining of corpora *about* scien-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

tific models than formulating (digital representations of) models in the first place.

In short, the quantitative techniques derived from conceptual space notions of concepts' dimensional overlap or prototyping are more relevant for mining data sets already established than for constructing and sharing data sets in the first place. These are the same quantitative methods which we earlier qualified vis-à-vis applicability to natural language: we argued that there are many cases of concept-combination that do *not* involve (at least to a substantial effect) numerically analyzable dimensional blends, and that role-based situational models can take the place of quantitative inter-dimensional calculations. It seems plausible that role-based models could offer a similar perspective on concept-blending in the area of data models.

When formulating a data model, the core project is to identify the data types through which information contained in (instances of) the model may be classified, to define those data types in terms of the specific data points and fields they aggregate. The key organizing principle is data *fields*, which aggregate into data structures, or compound values, that are type-instances (which therefore have internal structure, but also some integrity as a single conceptual unit). Data fields can in turn be single or multi-valued, with multi-valued fields typically being variant-sized collections, such as lists or unordered sets. A specific data point is attached to a type-instance through the value of an individual (single) data field or one value inside a collection (multi-value) field. Via such connections, each data point contributes some information to the total data encompassed by the larger type-instance. In effect, type-instances encapsulate the totality of information content contained in their component parts.

None of this discussion is particularly original or insightful, but it is worth highlighting these basic principles so as to draw attention to the implicit role of *information content*: each field within a type-instance amplifies the larger instance's total information content, which is the basic rationale for joining fields to instances. Data models are successful to the degree that they reinforce this phenomenon. For instance, metadata associated with specific data fields is valuable if it *increases the degree to which* the field-value augments the information content of the larger object. Annotating fields with scale and metric info (referring back to Chapter 6, Section 6.2.1) provides benefit, because, at least in many contexts, such annotations increase the amount of information supplied by an individual data field.

This perspective points to a strategy for integrating conceptual space theory more rigorously with data-modeling paradigms: we can adopt conceptual spaces to study the degree and nature of information-content elevation, which is associated with individual data fields seen as information-contributors, whose contributions are in turn mediated (and potentially augmented) by the relevant data model. Conceptual space theory, as we have argued, tends to analyze *increase in information content* via quantitative models of conceptual-blends, but we have suggested that such quantitative accounts are merely one way to operationalize the basic idea of studying information-amplification as a construction-driven phenomenon. In the data-modeling context, we can similarly focus on the more general rise in information-content, which may or may not be driven by numerically analyzable combinations between data-contributing elements. What we can focus on instead is the core principle that data *models* add value to data *sets* by *increasing the degree to which data-contributing elements in each model-instance augment the information content of the overall model* (for the sake of discussion, we can call these *information delta* effects). When considering strategies for developing or refining data models, then, the essential question to ask is *how does a particular strategy contribute to the models' effect of raising the information-amplification of individual*

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

elements? This points to some tactics for defining, assessing, and formalizing data-modeling strategies: to rigorously characterize a strategy is effectively to clarify how that strategy contributes to information-delta effects.

Likewise, defining delta effects can lay a foundation for strategy implementations, in terms of kernel operations for query-evaluation virtual machines, for example, which would allow data-validation, and data-integration strategies to be implemented as sequences of kernel operations. That is to say, a theory of delta-effects should guide construction of virtual machine operation-sets in the data set and data-query contexts. In this sense, modeling information-delta vis-à-vis data sets serves as a continuation on the related notion of information-delta in the context of procedural code models discussed in Chapter 6: in both cases information-delta analysis can be practically operationalized via query-evaluation virtual machines.

9.4 Delta roles and conceptual space markup language

Recall that in Chapter 6, we highlighted "selection" and "instantiation" concerns in the context of queries against data sets. To reiterate, imagine a data set (or alternatively a database) as encoding the totality of its information in the form of a single graph, so that pulling information from the data set is analogous to performing graph queries. Assuming the data set is strongly typed, one property of such a graph is that all of its information content is sorted into type-instances. In general, this means that structures within the graph, which may be hypernodes, edges, properties, and so forth, supply data-points that define a specific type-instance (either the values of data-field indexed by name, or one value in a multi-value collection).

Data spaces can of course have many different structures; only in special circumstances will they be *explicitly* encoded via graphs. With a suitably general and expressive graph model, however—for instance, one which combines property and hypergraphs (as discussed in earlier chapters)—we can always treat existing data-set layout as isomorphic to a graph schema subject to certain evolutionary and usage/query constraints. For this reason we will proceed by anchoring all discussion in hypothetical data sets presented as hypergraphs, setting aside the details of how to translate queries formulated in other context to queries against hypergraphs.

9.4.1 Information delta and data modeling

Starting from the principle that certain graph-sites anchor type-initialization opportunities, we can refer to an *instantiation neighborhood* of a graph, demarcated by all parts of the neighborhood supplying data to the same type-instances. For example, a type-instance may be encoded via one hypernode, leveraging values inside the hypernode, along with (potentially) properties asserted on the hypernode itself or any node inside it (which may carry other data-points). We can then expand the neighborhood to include other hypernodes, which are necessary to initialize the neighborhood's "root" instance. In this sense neighborhoods may overlap.[14]

According to the terminological conventions, we employ here, a "site" in a graph is any structuring element: hypernodes, "hyponodes" (nodes inside other nodes), edges, properties, channels, named subgraphs, and so forth. Sites are associated with type-instances based on the

[14] If desired, we can distinguish hypernodes whose purpose is to provide the equivalent of individual data fields—e.g., multi-value collections—from top-level hypernodes that carry instances of types which are primary from the point of view of the governing data model; secondary type-instances are then logically "part of" primary type-instances, providing at least in terms of logical interpretation a "nesting" effect characteristic of hypergraphs.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

neighborhood where they are contained.[15] Each site then contributes some data, directly or indirectly, to its associated type-instance. Any query engine targeting the graph must therefore be able to identify *how* sites contribute data, and incorporate these details into its query-evaluation strategies, particularly in the context of selection and instantiation queries.

To demonstrate, first consider "instantiation" or "initialization" queries, whose recall are those confirming that in the neighborhood of a given graph-site we have sufficient data to populate an instance of some type. Essentially, this means at least that there are data points for every field that is required for initializing a value of that type, where these data points are embodied in structures attached to the site. Determining instantiation-queries therefore involves matching requisite fields against sites in the relevant neighborhood providing those fields' data. For such a process to proceed unambiguously, neighborhood-sites would be annotated with metadata specifying how they contribute to instance-initialization for some type; this metadata may be inferred by the graph engine or directly asserted by the governing data model. If it is inferred, we assume that the basis for this inference is some unambiguous property of the data model, with the data modeling language being formulated at least in part to drive that kind of inference. For instance, associating subvalue-nodes (hyponodes) with a named field and declaring that field as necessary for type-initialization—as part of the declaration for some type—signals that a subvalue indexed via that name is an initialization-precondition, and that the relevant hyponode thereby plays an initializing role in the surrounding neighborhood. The relevant type- and field-documentation therefore serves as a form of initialization-annotation.

[15] Without further qualification we will use "neighborhood" in this discussion to loosely mean "instantiation" neighborhoods (in our terms; note that these are somewhat different now from the definitions we proposed in Chapter 6).

In effect, then, we can stipulate, as a precondition on a suitable data-modeling framework, that annotations may define how data sites contribute to type-instantiation (allowing for this metadata to be implicit in other related declarations). Considered from the perspective of *information content*, we can then investigate how each site's contributions to an initialization (the nature of the information supply) augment information content in different ways.

For example, suppose we have either a property or a subvalue-field providing a named field (in a record-like data structure), and *moreover* an annotation asserting units of measurement. This could take the form of a global guarantee that values for this field will always be represented according to a specific measurement-scale, or alternatively a contract that such units-data is available as a component data-point in the scale-delimited values on a case-by-case basis. These annotations therefore provide information about units of measurement, which affects how the relevant values may be used to initialize neighborhood type-instances.

It is conceivable that scale-annotations are not necessary in some contexts because of alignment between data sources and data-processing routines, which ensures that measurement details are fixed and unambiguous for the lifetime of the data set, but in general the presence of scale-annotations would supply a greater quantity of information than raw values without such annotation. If, under these motivations, a data model explicitly requires that units declarations (for any quantities that are not simple magnitudes) be confirmed as a precondition for type-initialization, then that facet of information content supplied by the annotated site fits in to the type-instantiation concern; it is relevant to instantiation queries and forms part of the interface implicit in applicable data sets for constructing values of the corresponding type.

On the other hand, if units-annotations are not requisite for *instantiation*, they may still be important for *selection* queries, if ranges in the

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

corresponding data type are part of query criteria, which would filter potential type-instances. When (say) querying a bioimage database for diameters of regions-of-interest, it is necessary to be sure that one is comparing regions by the proper scale (e.g., centimeters, or percentage of image-width) against the query range.

In any case, the information provided by a given graph-site can be classified into different facets, which we may call *roles* (establishing a connection to multi-relations as discussed above), and these facets can become relevant for different varieties of queries. The *type* of data represented at a site, along with its actual raw value, correspond to two different roles: for initialization queries (if we only want to ascertain that initialization is possible in some neighborhood, which is different from *constructing* the type-instance) the *type* is almost certainly relevant, whereas raw values are not; but raw values would come into play for selection-related queries.

Metadata, such as units of measurement, could potentially be associated with type-declarations, as a kind of further contract annotating the type, and come into play for questions about whether instantiation is possible. Conversely, scale-declarations may not be directly examined until raw data is actually compared against some range. But in either case the data model expresses the *sorts* of queries where meta-data characterizing the site information's *role* becomes relevant. As for scale-delimited values, similar comments would apply to other meta-data properties associated with data fields, such as valid ranges, or expected distributions (quantifying how much a fixed value for the field is typical or atypical, to the degree that this can be asserted of a single field in isolation from others in its enclosing type), or such as similarity measures (quantifying the degree to which differences in value on some axis, or the lack thereof, contribute to dissimilarity or similarity between two type-instances).

Insofar as each site contributes to some *increase* in a data set's overall information content, these *roles* serve to define this "delta" effect more granularly. For this reason, we will refer to such roles more precisely as *delta* roles, with the idea that for each delta role there is a specific mechanism through which some data or meta-data adds information content: as a type attribution, a raw value, a scale (or range/distribution/distance-metric etc.) annotation, and so forth. Delta roles can then combine with assertions of query *contexts*, where the role-specific information comes into play (e.g., type-initialization vs. instance-selection).

We propose these meta-data annotations being implemented as explicit data-modeling features, rather than left implicit in database schema. These annotations could then be directly exploited by query-evalution virtual machines; they might be built in to the virtual machine architecture and operation-set to a degree that would be infeasible without an explicit accommodation for such metadata in the data-modeling paradigm. Fig. 9.1 outlines these ideas through (what we'll call) a "localized" syntagmatic graph, diagramming relations among sites in a hybrid property-hypergraph via a syntagmatic representation. Delta roles can also aid in traversal implementations; for example, checking treatments of scale-units (recall Chapter 6's discussion of tracing program flows relative to scale-sensitive procedure inputs/outputs) could proceed by following paths determined by graph-sites annotated with declarations that scale-unit meta-data is in effect.

A formal elaboration of role-delta information would overlap significantly with conceptual space implementations, particularly as these are applied to data-modeling (notably CSML); as such, we incorporate CSML into the query system considered here for data integration, specifically as a component of the mechanism for defining and using delta-roles. We will explain this tactic

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

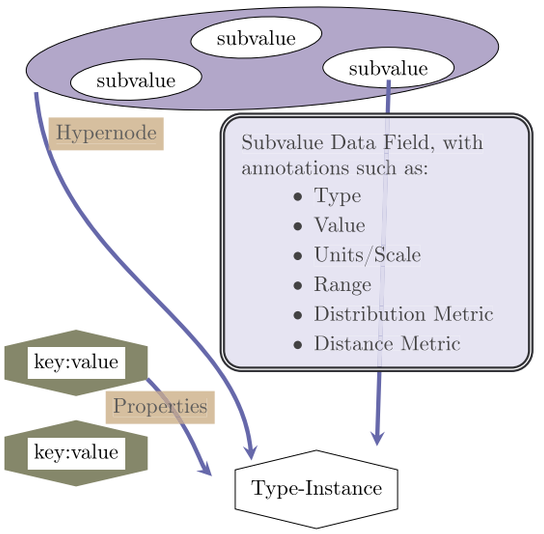

FIGURE 9.1 Role-deltas on a localized syntagmatic graph.

for incorporating CSML with reference to sample data-sets published in this book's supplemental materials.

9.4.2 The artificiality of data semantics

One challenge facing any theory which attempts to mix ideas from both (on the one hand) natural-language semantics and (on the other) contexts such as data modeling and/or programming languages is that semantics in the former sense exists in cognitive and conceptual surroundings, which do exist in the latter sense. Natural language is spoken by people, in (typically) enactive, interpersonal, goal-directed circumstances, where the ambient situation provides frames of reference that prime certain conceptual foregrounds and backgrounds. Language is not suspended in a pure logical space, but rather supplements each person's prelinguistic comportment to the scenarios wherein linguistic activity takes place.

This is one sense (which we will not explore further here, but would be worthy of detailed study in a more philosophical context) that linguistic meaning should be conceived in "delta" terms, i.e., in terms of information-increase: any meaning expressed in language *adds* content to the totality of conversants' cognitions already believed, perceived, experienced, or posited as an object of thought. Many of the structures we might use to characterize languages—syntagmatic constructions involving thematic roles, situational context, and speaker-relative epistemics, for example—also apply to cognitive attitudes, which are equally present in the absence of (at least explicit) language (allowing for the possibility that language-acquisition *in general* restructures cognitive dispositions in ways that leave an impression in how people perceive the world around them, so that language is always somehow virtually present; we speak here only of *concrete* linguistic content). Any linguistic entity or pattern need merely trigger prelinguistic situation-comprehension faculties to play its relevant constructional role.

None of this prelinguistic background is especially relevant for the semantics of digital data and computer code, except for AI projects, which try to bridge the gap between *in silico* calculations and human cognitions. The issues of symbol-grounding and construction semantics, which we touched on at the beginning of this chapter, are preconditions for a realistic AI simulation of language competence, and embodies lines of research connecting conceptual space theory to computational models. In this section, however, we are concerned with more prosaic data and code modeling scenarios. Since in such contexts there is no ambient prelinguistic consciousness that can ground semantics in cognitive (situational/perceptual) dispositions, a minimal requirement for discussing data/code semantics is what such semantics *is*. Articulating what exactly falls under the scope of "semantics" in these artificially formal contexts helps to

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

clarify the intuitions behind particular authors' semantic models.

We have already defined the basic elements of our specific construal of data/code semantics, so at this point we need merely connect them to this underlying topic. Insofar as *data semantics* involves the useful structuring, encoding, serialization, and interface-design for data sets and data bases—particularly data sets or database schema that can be annotated with metadata clarifying the semantic constraints respected in the overall data-curation process—the core elements of data semantics are embodied in *instantiation queries* and *selection queries* (here we will define "initialization" queries as a follow-up to instantiation queries, thus establishing the technical distinction between these terms). To reiterate (see Section 7.1 for our preliminary discussion). Instantiation queries are, then, concerned with whether a data-type instance can be initialized from a given site in a data set; initialization queries are those used to populate data fields necessary to actually perform such instantiation; and selection queries ascertain whether a type-instance would meet given criteria (were it to be instantiated). Data semantics should also recognize the possibility of "fiat" instantiation, where queries against type-instances can be evaluated without actually performing an instantiation-step if there is some other way to pull the relevant information from a data set/base.

Alongside data-semantics, whose key phenomena are "queries" of the forms just mentioned—or algorithms/capabilities which are functionally similar to such queries—we would also emphasize *code semantics*, whose essential elements are mapping symbols (or graph-nodes) to procedures (procedure resolution), and dynamically calling or examining procedures with parameters populated from data-structure (e.g., graph) sites (call this *procedure reflection*). In this context, we will allow the term *reflection* to include static as well as dynamic analysis of procedure metadata, such as finding pre- and post-conditions or selecting which is the best procedure to call from a group of plausible candidates. Such reflective capabilities would certainly apply to the process of resolving runtime-reflection calls, but we will more generally speak of "reflection" as any process for acquiring and/or incorporating procedural metadata at one or more stages of software engineering, including compilation and runtime as well as pre-compilation static analysis, requirements engineering, design and testing, and so on.

In any case, considering both data and code, the core elements of data/code semantics together are *instantiation queries*, *initialization queries*, *selection queries*, *procedure resolution*, and *procedure reflection*. We claim that almost all "semantic" issues in computational contexts (again excluding NLP-inflected AI) can be classified into one of these five concerns. This assertion, to the degree that it holds up, can propagate to the design of (code and data) annotation schemes and query expression/evaluation systems. For each element of an annotation or query expression, we should be able to identify which of the five core code/data semantic concerns it targets. Likewise, query evaluation systems could incorporate this code/data semantic model into their implementation architectures: primitive query-engine operations can be grouped according to which of the five concerns *they* target. We propose to employ this overall architecture to formulate a query and serialization language, as well as a query-engine virtual machine, which would incorporate precedent technologies such as conceptual space markup language and the text-as-graph markup language (TagML) [7], [6], [18].

9.5 Conclusion: Toward a scientific data semantics

Subsequent to conceptual space theory being applied to technology—since its stepping outside its philosophical/linguistic origins—one of

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:57 • p. 29
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

the central goals has been to develop a more useful and realistic semantic model for data structures (more or less what we here refer to as "data semantics"). This is certainly evident in explicit comparisons against the core technologies of the semantic web—ontologies, resource description format (RDF), and so forth—and the outright argument that "the Semantic Web is not very semantic" [24, page 2]. Here Gärdenfors implicitly critiques the RDF framework and ontologies in general (there is nothing in this argumentation specific to science), although projects such as CSML point to conceptual spaces being particularly emphasized as semantic models for *scientific* data. We too have focused in this direction, though we would contend that a rigorous semantics for scientific data could potentially be generalized to shared/networked information in many domains, akin to a semantic web. In this concluding section, though, we intend to consolidate our discussion focusing again on scientific data models.

It could be argued that the process of sharing scientific data does not actually require a data *semantics*. Most scientific data is in a tabular, vector, or matrix form, which (taken out of context) is just a jumble of numbers; it is only when plugged in to the appropriate software, or evaluated in its intended theoretical context, that these numbers actually become part of "science." In this case, one might argue, there is no particular reason to express scientific contexts within data sets themselves; data sets should simply record the minimum information needed to send or reproduce data across different research environments.

Contrary to that assessment, however, there *have* been numerous efforts in the research/academic community to codify scientific data models (some of which we listed in Section 2.2). At least some scientists, in short, have argued that published scientific data should be organized and annotated in a manner which documents scientific assumptions and contexts, rather than just serializing raw numbers. To the degree that researchers seek to formulate *expressive* data sets along these lines, then issues of "scientific data semantics" come to the fore. Exploring semantic theories to support such a data-semantics can draw in considerations from both science and digital technology, and even from linguistics and philosophy.

With that said, there is only a limited amount of information that can be provided via "static" structures within which data is encoded. A lot of scientific information (like information in general) is naturally expressed as *records*, tuples of individual fields with their own names or labels (**PatientName** and so forth), so at the very least field-names provide a conceptual overview of data semantics: indeed, this is the primary source of data-integration architectures within broad-based biomedical projects, such as OMOP or CDISC (see Chapter 2). Controlled vocabularies attempt to render semantics based (primarily) on field-names (or, analogously, on column-names in the case of tabular data) more rigorous by ensuring that multiple parties use the same names or labels for conceptually similar units of information. Semantic web ontologies also build off of field-names by stipulating common axioms constraining information culled from different sources, which utilize restricted field-names derived from controlled vocabularies; in this case, not only the textual *name* of the field, but also certain structural contracts in the data associated with the field (for instance, that a given field, e.g., patients' *first* name, is always paired with other fields, e.g., *last* name) is aligned between disparate data-providers.

Relatively "static" data models can also define and distinguish one-to-one, one-to-many, many-to-one, and many-to-many relationships (viz., different forms of relation cardinality), which provides another source of general conceptual overviews of data structures. For example, a patient presumably has only one (full) name, but they may be taking multiple medications. Likewise, pharmaceuticals may have only

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

one chemical formula, but they may be taken by many patients (in a clinical study, for example).

Yet (as we argued above at the end of Section 9.3.1) one can achieve only a limited degree of semantic precision via "static" details about (or meta-models constraining) data structures, such as field-names or relation-cardinality. This appears to be the gist of conceptual space arguments against the semantic web, and as such motivates proposals for more detailed statistical annotations, addressing issues such as units of measurement, valid ranges, the fusion of *dimensions* into *domains*, and similar meta-modeling constructions. Still, these are essentially static forms of meta-data, even if they lend greater theoretical precision than just "raw numbers."

Over and above static meta-data, we contend that a well-motivated data semantics will be predominantly "procedural," by which we mean that the actual "semantics" of scientific data—the empirical, theoretically informed meanings or interpretations one assigns to numeric quantities or other information-values measured or observed as part of a scientific investigation—depends on computations, where those values are analyzed. Any static meta-data, from field-names to dimensional annotations, can provide only a summarial precis of research data's scientific meaning. The full-fledged scientific significance of a given data structure depends on the theory and investigations where it originates, and to the degree that such research has a digital residue, it would lie in the set of procedures (algorithms, calculations, and so forth) that project a scientific model into the computational domain. In this sense, it is unrealistic to propose a *semantics* for scientific data, *except* in the context of procedure-collections (e.g., code libraries) that manipulate such data.

This hypothesis has consequences from both theoretical and practical perspectives. Practically speaking, one entailment that seems to follow, a conclusion we would certainly advocate, is that code reuse is an intrinsic part of data sharing. It is now common practice for scientists to deposit raw research data in a public archive, and while such transparency and data-availability are preferable to the alternative (where data is not published at all), scientists should be encouraged to develop their data sets as "research objects" or similar "FAIR" resources, where data and code are bundled together. This adds a burden to the publishing process, which should not be underestimated: preparing a reusable code base could easily become the most time-consuming part of a research project. But scientists can turn to reusable code as a forum for demonstrating their theories and methods in an interactive, data-driven fashion. Moreover, we will argue that data-set implementations can facilitate scientific projects, even prior to the stage of open data sharing.

9.5.1 Research data and data integration

In this book we have employed the summarial figure of a "semiotic saltire" to describe a typical pattern in the *organization* of procedures within a multi-faceted code base (e.g., integrated software components, such as "multi-aspect modules," returning to terminology from Chapter 4). To the degree that code accompanying scientific data covers multiple software-engineering concerns, it is likely that the resulting procedures will end up grouped into aspects according to a pattern similar to the saltire, which in this case can potentially serve as a rough guide to identifying coding requirements. (That is, we intend the saltire model to be partly normative—this is how modules often *should* be organized—but also partly observational, i.e., procedures *tend* to group into such a pattern.) When preparing the code base for research data, concerns identified in the saltire tend to be foregrounded: how should the published data appear in GUIs? How should it be derealized and deserialized? How should it be structured for database persistence?

Though we assume that such a "structuring" of requirements applies to shared research

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:57 • p. 31
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

data, we would give similar analyses for context such as clinical research networks or multi-site clinical trials. During a trial's planning stages, for example, investigators might benefit from modeling the information generated during the course of the trial as a *de facto* research data set (of course, when trial results then become published academically, some of that data will in fact *be* released as research findings). Data-set semantics applies more generally than just in the context of research results curated as open-access data sets; conceiving clinical-trial data as a still-emerging *research* data set can help structure the programming and data-collection logistics governing how the trial will digitally operate, and how (assuming a multi-site project) information from different institutions will be aggregated.

Insofar as research and/or clinical data are curated according to the norms of featureful open-access data sets (e.g., research objects), and if the resulting code base accepts the basic premise of multi-aspect design, then the resulting information resources will have from the outset a programming environment equipped with a useful variety of software capabilities: custom GUI classes, implemented protocols for data (de)serializing/marshaling, and so forth. Individual clinical trials, for example, can be encapsulated in distinct *modules*, which could be injected into clinical applications. All of this structure might then be leveraged for scenarios such as machine-learning or data-mining/integration.

For example, consider the task of unifying results from two different multi-site clinical trials. If each trial's data comes packaged in self-contained modules spanning multiple programming aspects, developers would have a valuable body of code already implemented for each information-space, which is more convenient than needing to write code *do novo* when presented with research or trial results as raw data. As a concrete example, suppose one stage of data integration requires the use of GUI components for human users to provide feedback about how two distinct data sets should be merged. If GUI classes are provided as part of the original data sets' modules, they would not have to be engineered from scratch within a data-integration context.

Aside from such practical maxims, however, this book has also focused on data semantics from a natural-language perspective, particularly that of conceptual spaces. How can we draw intuitions from natural language in the context of a predominantly *procedural* semantics? If the interpretive substance behind any scientific data only emerges in the context of procedures where that data is manipulated and analyzed, it would seem difficult to press into service any "static" meta-model (whether based on conceptual spaces, on web ontologies, or anything else), which would be logically removed from procedures themselves. The only real "semantics" applicable to a given scientific data-space would need to be assessed by looking at the specific procedures implemented for that data and how they exemplify the relevant scientific model/theory, through algorithms and data constructors.

In short, our discussion leads to the problem of formulating a *procedural* conceptual space semantics, to the degree that we wish to sustain the intuition of conceptual space theory as a richer and more realistic semantics alternative to (say) the canonical semantic web. We will therefore conclude by sketching a few ideas in this context.

9.5.2 Toward a procedural conceptual-space semantics

The essential insight of "procedural" semantics, at least in the context of scientific data, is that meanings and interpretations of scientific measurements/observations are dependent on the specific scientific theories guiding research, where the data originates, and these are only computationally manifested in any substantial

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:57 • p. 32
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

way through procedures (algorithms, analyses, visualizations, and so forth). There is not a lot of semantic detail that can be provided by overarching computing environments *apart from* procedures implemented for each specific domain. A programming environment may provide tools to *facilitate* implementations, but in the absence of procedures, actually composed in concrete fashion (expressing theoretical axioms, calculations, or algorithms via source code) we cannot attribute a significant *semantics* to such programming environments. The following question then becomes relevant: granted that the semantic weight of a data-model rests predominantly on procedures formulated for an associated code model, how can the programming environment and computational tools, which enable those procedures to be implemented, at least *reinforce* the semantic details encapsulated via procedures themselves?

To examine this question, we'll point out first of all that procedures tend to cluster into interrelated groups. Moreover, this clustering effect is often correlated with data structures insofar as they would be represented within a data model. Consider the general case of multi-valued data fields (i.e., one-to-many relationships), such as the list of a patient's medications. Multi-valued collections require several different procedures to be fully manipulated, at the minimum, those for *inserting* and for *removing* values. So (reprising Section 6.2.2) code managing a patient's health records, for example, might include a procedure, which has the effect of *adding* reference to a certain medicine to the list, which the patient is currently taking, and a second procedure for *removing* a medication from that list. These two procedures, of course, are logically inter-related. This is a simple example of how meta-modeling paradigms often propagate to *inter-procedural* relations, a tendency we discussed in Chapter 6 in the context of using code models to *instantiate* data models.

Chapter 6 also discussed data-modeling techniques, such as scale/dimension annotations or remote procedure calls (i.e., modeling the preparatory code needed to expose procedures, or capabilities dependent on specific sequences of procedures, as an external service; we proposed the term "meta-procedure"). These likewise furnish examples of how *networks* of procedures concretize data-model paradigms. For example, validating scientific scales and units of measurements may involve preparatory code, which ensures dimensional alignment as a precursor to performing some specific calculation, leading to functionality being split between two contexts: a "preparatory" procedure which performs a function we might refer to as "gatekeeping" [13], and then the "primary" procedure which enacts the requisite computations. Here the logic governing the relation between the "gatekeeping" and "primary" procedures manifests data-modeling concerns (in this case those of dimensional analysis and consistency), analogous to the case of multi-valued data fields being supported by both *insertion* and *deletion* procedures.

Given these implicit logical connections between procedures, programmers have the option of *explicating* such connections via code annotations or other interface-description techniques. Insofar as procedure-interrelationships concretize and originate from data-modeling concerns, notating procedural connections likewise serves the goal of transparently describing data models operating in the context of the current code base. In short, tools for constructing and identifying procedural annotations, particularly insofar as these allow procedure's clustering patterns to be described and rationalized, serve as one technology for elucidating data semantics through code components/modules.

There are multiple criteria that could be applied when modeling how procedures within a given library or component are interconnected. One can trace dynamic *execution flows* in the sense of identifying, for a given run of a pro-

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:57 • p. 33
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •

B978-0-32-385197-8.00016-7, 00009

gram/application, which procedures are called prior or subsequent to which others. We can also consider (more generally) which procedures *might* be called prior to others. Of course, one straightforward inter-procedural relation is when one procedure calls a second. Alternatively, an enclosing procedure might call an antecedent, and then subsequent procedure in sequence. These cases are distinguished in terms of whether the prior procedure returns before the later one begins/resumes. A sufficiently expressive pointcut expression language (as we explored in Section 6.2) can identify locations in source-code, where specific kinds of inter-procedure relations are exercised (one calling a second, one being called before a second in an overarching procedure, and so forth).

In and of itself, such program-flow connections do not necessarily have a specific "semantic" interpretation; they may merely reflect operational sequences as specific algorithms or implementation patterns are encoded, but at least on some occasions there is a meaningful semantics behind inter-procedural connections manifest at the program-sequence level (an example would be "gatekeeping" checks as mentioned earlier; another would be deserializing input data structures, a preparation for handling a remote meta-procedure invocation). Insofar as pointcut expressions can single out code sites, which *do* have such semantically substantial rationales, the use of pointcuts to construct rigorous code models can provide a technique for clarifying how data semantics are manifest within a code base, establishing the interactions between data and code models, which we discussed in Chapter 6.

We might envision code libraries/modules as "procedural" spaces, whose underlying structures are constituted by semantically meaningful inter-procedural relations that can be annotated and described. A *procedural* space is, of course, not the same thing as a "conceptual" space, but this chapter has reviewed how conceptual space semantics are often formalized by considering the mutation in information content as one moves between sites of "transformations" that may be seen as analogous to procedures; procedures themselves in the context of computer code, or "morphisms" in hypergraph categories, and verbs in natural languages. In other words, a conceptual space semantics often emerges from networked procedure-spaces, or representations structurally akin to them.

Our discussion in this chapter has attempted to highlight one potential avenue for deriving a rigorous conceptual space semantics in a procedure-network context (whether this is defined explicitly or implicitly) through the lens of information-content "amplification" and "delta" paths/roles. We suggest that this is a promising avenue for future research, even if our analysis to this point only presents the initial step to a theory along such lines. Probably the trajectory of such a theory's development cannot be driven by abstract concerns alone, but rather in a feedback circle informed by specific scientific data sets, for which data-semantics can be assessed concretely, and through the implementation of code-annotation systems, where inter-procedural connections can be notated and analyzed.

References

[1] Benjamin Adams, Martin Raubal, A metric conceptual space algebra, in: International Conference on Spatial Information Theory, 2009, pp. 51–68, https://pdfs.semanticscholar.org/521a/cbab9658df27acd9f40bba2b9445f75d681c.pdf.

[2] Jens Allwood, Semantics as meaning determination with semantic-epistemic operations, in: Jens Allwood, Peter Gärdenfors (Eds.), Cognitive Semantics: Meaning and Cognition, John Benjamins, 1999, pp. 12–28, https://benjamins.com/catalog/pbns.55.02all.

[3] Lucas Bechberger, Kai-Uwe Kühnberger, A comprehensive implementation of conceptual spaces, in: Artificial Intelligence and Cognition, Proceedings, 2017, http://ceur-ws.org/Vol-2090/paper4.pdf.

[4] Lucas Bechberger, Elektra Kypridemo, Mapping images to psychological similarity spaces using neural networks, in: Artificial Intelligence and Cognition, Pro-

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

ceedings, 2018, http://ceur-ws.org/Vol-2418/paper3.pdf.

[5] Ismaïl Biskri, Jean-Pierre Descles, Applicative and combinatory categorial grammar (from syntax to functional semantics), in: Ruslan Mitkov, Nicolas Nicolov (Eds.), Recent Advances in Natural Language Processing, John Benjamins, 1997, https://www.researchgate.net/publication/232754402_Applicative_and_Combinatory_Categorial_Grammar_from_syntax_to_functional_semantics.

[6] Elli Bleeker, et al., Agree to disagree: modelling co-existing scholarly perspectives on literary text, https://academic.oup.com/dsh/article-abstract/34/4/844/5576174?redirectedFrom=fulltext.

[7] Elli Bleeker, et al., Between flexibility and universality: combining TAGML and XML to enhance the modeling of cultural heritage text, http://ceur-ws.org/Vol-2723/short39.pdf.

[8] Greg Carlson, Thematic roles and the individuation of events, https://www.sas.rochester.edu/lin/people/faculty/carlson_greg/assets/pdf/them-roles-events.pdf.

[9] Stuart Chalk, SciData: a data model and ontology for semantic representation of scientific data, Journal of Cheminformatics 8 (2016), https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0168-9.

[10] Lucas Champollion, Covert distributivity in algebraic event semantics, Semantics & Pragmatics 9 (2016) 1–65, https://semprag.org/article/view/sp.9.15.

[11] Lucas Champollion, Overt distributivity in algebraic event semantics, Semantics & Pragmatics 9 (2016) 1–65, https://semprag.org/article/view/sp.9.16.

[12] Lucas Champollion, Parts of a Whole: Distributivity as a Bridge Between Aspect and Measurement, Dissertation, University of Pennsylvania, 2010, https://repository.upenn.edu/cgi/viewcontent.cgi?article=2117&context=edissertations.

[13] Nathaniel Christen, Hypergraph type theory for specifications-conformant code and generalized lambda calculus, in: Amy Neustein (Ed.), Advances in Ubiquitous Computing: Cyber-Physical Systems, Smart Cities, and Ecological Monitoring, Elsevier, 2019.

[14] Bob Coecke, et al., Interacting conceptual spaces I: grammatical composition of concepts, Extended version of Proceedings of the 2016 Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science, pp. 11–19, https://arxiv.org/pdf/1703.08314.pdf.

[15] Bridget Copley, Force dynamics, in: Robert Truswell (Ed.), Oxford Handbook of Event Structure, Oxford, 2019, pp. 103–149, http://bcopley.com/pubs/force-dynamics.

[16] Ann Copestake, et al., An architecture for language processing for scientific texts, in: UK e-Science Programme All Hands Meeting, Proceedings, 2006, https://abdn.pure.elsevier.com/en/publications/an-architecture-for-language-processing-for-scientific-texts.

[17] Suelen M. de Paula, Ricardo R. Gudwin, Evolving conceptual spaces for symbol grounding in language games, Biologically Inspired Cognitive Architectures 14 (2015) 73–85, https://www.sciencedirect.com/science/article/pii/S2212683X15000493.

[18] Ronald Haentjens Dekker, et al., Parsing a markup language that supports overlap and discontinuity, in: Document Engineering, Proceedings, 2020, pp. 1–4, https://dl.acm.org/doi/abs/10.1145/3395027.3419590.

[19] Stefan Dietze, John Domingue, Exploiting conceptual spaces for ontology integration, in: 3rd Asian Semantic Web Conference, 2008, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.4434&rep=rep1&type=pdf.

[20] Gilles Fauconnier, Mental Spaces: Aspects of Meaning Construction in Natural Language, Cambridge University Press, Cambridge, 1994.

[21] Gilles Fauconnier, Mark Turner, The Way We Think: Conceptual Blending and the Mind's Hidden Complexities, Basic Books, 2002.

[22] Johannes C. Flieger, Gradable Adjectives and the Semantics of Locatives, Dissertation, University of Edinburgh, 2009, https://era.ed.ac.uk/bitstream/handle/1842/3995/Flieger2009.pdf?sequence=1&isAllowed=y.

[23] Peter Gärdenfors, Does semantics need reality?, in: Alexander Riegler, et al. (Eds.), Understanding Representation in the Cognitive Sciences, Springer, Boston, 1999, pp. 209–217.

[24] Peter Gärdenfors, How to make the semantic web more semantic, https://slab.org/tmp/Gardenfors04.pdf.

[25] Peter Gärdenfors, Primary cognitive categories are determined by their invariances, Frontiers in Psychology (2020), https://www.frontiersin.org/articles/10.3389/fpsyg.2020.584017/full.

[26] Peter Gärdenfors, Massimo Warglien, Using conceptual spaces to model actions and events, Journal of Semantics 29 (4) (2012) 487–519, https://www.researchgate.net/publication/274999478_Using_Conceptual_Spaces_to_Model_Actions_and_Events.

[27] Gabriel Gaudreault, Derivational Event Semantics for Pregroup Grammars, Dissertation, Concordia University, Montreal, 2016, https://spectrum.library.concordia.ca/981873/9/Gaudreault_MA_F2016.pdf.

[28] T. Florian Jaeger, Redundancy and reduction: speakers manage syntactic information density, Cognitive Psychology 61 (1) (2010) 23–62, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896231/.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

[29] Ronald Langacker, Cognitive Grammar: A Basic Introduction, Oxford University Press, Oxford, 2008.
[30] Ronald Langacker, Interactive cognition: toward a unified account of structure, processing, and discourse, International Journal of Cognitive Linguistics 3 (2) (2014), http://lchc.ucsd.edu/MCA/Mail/xmcamail.2014-08.dir/pdf8jEY9UCVVh.pdf.
[31] Ronald Langacker, The English present: temporal coincidence vs. epistemic immediacy, in: Adeline Patard, Frank Brisard (Eds.), Cognitive Approaches to Tense, Aspect, and Epistemic Modality, John Benjamins, 2011, https://benjamins.com/catalog/hcp.29.06lan.
[32] Beth Levin, Malka Rappaport Hovav, Lexicalized meaning and manner/result complementarity, in: B. Arsenijević, et al. (Eds.), Studies in the Composition and Decomposition of Event Predicates, Springer, 2013, pp. 49–70, https://web.stanford.edu/~bclevin/barcel11rev.pdf.
[33] Ian Lewin, Using hand-crafted rules and machine learning to infer SciXML document structure, https://www.semanticscholar.org/paper/Using-hand-crafted-rules-and-machine-learning-to-Lewin/3dd2756e42ae24df0769711c3b7a55249d7b17cf.
[34] Peter McQuilton, et al., BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences, Database 2016 (2016), https://pdfs.semanticscholar.org/504a/e934c99ed5cbd4ea925ec13ebf86f553e251.pdf.
[35] Carlos Montemayor, et al., Implementation, formalization and representation: challenges for the integrated information theory, Journal of Consciousness Studies 26 (1–2) (2019) 107–132, http://online.sfsu.edu/barros/publications/publications/files/MontemayorEtAl2019.pdf.
[36] Erwan Moreau, From link grammars to categorial grammars, in: Proceedings of Categorial Grammars 2004, 2004, pp. 31–45, https://hal.archives-ouvertes.fr/hal-00487053/document.
[37] Tiago Nunes, et al., BeCAS: biomedical concept recognition services and visualization, https://www.ncbi.nlm.nih.gov/pubmed/23736528.
[38] Matías Osta Vélez, Inference and the Structure of Concepts, Dissertation, Ludwig-Maximilians-University, 2020, https://edoc.ub.uni-muenchen.de/27633/7/Osta_Velez_Matias.pdf.
[39] Joey Pollack, Holism, conceptual role, and conceptual similarity, https://www.tandfonline.com/doi/abs/10.1080/09515089.2020.1729973?journalCode=cphp20.
[40] Dietrich Rebholz-Schuhman, et al., IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules, https://www.semanticscholar.org/paper/IeXML%3A-towards-an-annotation-framework-for-semantic-Rebholz-Schuhmann-Kirsch/1d72a56b6576117c62f388a5f2193965e4c7e293.
[41] John T. Rickard, A concept geometry for conceptual spaces, Fuzzy Optimization and Decision Making 5 (2006) 311–329, https://altexploit.files.wordpress.com/2017/06/a-concept-geometry-for-conceptual-spaces.pdf.
[42] Bradley Rives, The empirical case against analyticity: two options for concept pragmatists, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.9556&rep=rep1&type=pdf.
[43] Aurélie Rossi, Applicative and combinatory categorial grammar: analysis of the French interrogative sentences, in: FLAIRS Conference, Association for the Advancement of Artificial Intelligence, 2008, https://www.aaai.org/Papers/FLAIRS/2008/FLAIRS08-118.pdf.
[44] C.J. Rupp, et al., Flexible interfaces in the application of language technology to an eScience corpus, https://www.cl.cam.ac.uk/~sht25/papers/Rupp_et_al.pdf.
[45] Dan Ryder, Problems of representation II: naturalizing content, in: The Routledge Companion to Philosophy of Psychology, Routledge, 2009, pp. 251–279, https://www.semanticscholar.org/paper/Problems-of-representation-II%3A-naturalizing-content-Ryder/1073bf288423e3e9b1940628cb2deb7db8b8fae2.
[46] Bahar Sateli, René Witte, Semantic representation of scientific literature: bringing claims, contributions and named entities onto the linked open data cloud, PeerJ Computer Science (2015), https://peerj.com/articles/cs-37.pdf.
[47] Pedro Sernadela, José Luís Oliveira, A semantic-based workflow for biomedical literature annotation, Database 2017 (2017), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5691355/.
[48] Frank Andreas Sposito, What do data curators care about? Data quality, user trust, and the data reuse plan, http://library.ifla.org/1797/1/S06-2017-sposito-en.pdf.
[49] Gregor Strle, Semantics Within: the Representation of Meaning Through Conceptual Spaces, Dissertation, Novi Gorici, 2012, http://www.ung.si/~library/doktorati/interkulturni/25Strle.pdf.
[50] Anna Maria Tammaro, et al., Data curator's roles and responsibilities: an international perspective, Libri 69 (2) (2019), https://www.degruyter.com/document/doi/10.1515/libri-2018-0090/html.
[51] Miriam Taverniers, Subjecthood and the notion of instantiation, https://semanticsarchive.net/Archive/DU0OGRkO/Taverniers-2005-Subjecthood-PREPRINT.pdf.
[52] Chris F. Taylor, et al., Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project, Nature Biotechnology 26 (2008) 889–896, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771753/.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

[53] Simone Teufel, Min-Yen Kan, Robust argumentative zoning for sensemaking in scholarly documents, in: R. Bernardi, et al. (Eds.), NLP4DL/AT4DL, in: Advanced Language Technologies for Digital Libraries, 2009, pp. 154–170, http://people.cs.pitt.edu/~litman/courses/nus062615/readings/TeufelKan.pdf.

[54] Simone Teufel, et al., An annotation scheme for discourse-level argumentation in research articles, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 110–117, https://www.aclweb.org/anthology/E99-1015/.

[55] Ralf Vogel, Polyvalent Verbs, Dissertation, Humbolt University, 1998, https://edoc.hu-berlin.de/bitstream/handle/18452/15160/Vogel.pdf?sequence=1.

[56] Massimo Warglien, Peter Gärdenfors, Matthijs Westera, Event structure, conceptual spaces and the semantics of verbs, Theoretical Linguistics 37 (2012) 159–193, https://iris.unive.it/retrieve/handle/10278/37082/27662/semantics_of_verbs.pdf.

NEUSTEIN, 978-0-323-85197-8

B978-0-32-385197-8.00016-7, 00009

Non-Print Items

Abstract

This chapter will more substantially develop our approach to conceptual space theory in natural (as well as programming) language contexts, that was initiated in earlier chapters. We present further philosophical motivations for the structural details of our proposed "syntagmatic graph" representations and examine techniques for integrating conceptual spaces with linguistic paradigms, such as conceptual role semantics and situational semantics. Our central argument is that the classical linguistic concept of "thematic roles" provides an alternative framework for analyzing the semantic integration of multiple conceptual spaces, contrasted with "quantitative blend" models endemic to conceptual space theory proper. Therefore we propose "role-indexed" conceptual space models, which have distinct semantic and syntactic patterns, juxtaposing this theory to existing formalizations of conceptual spaces in (for example) Quantum NLP. With that natural-language foundation as a motivation, we then consider semantic models as they could be more concretely applied to scientific data sets.

Keywords

conceptual space theory, syntagmatic graph, role semantics, Quantum NLP, verb-centric grammar, dimensional analysis, emergent syntax, instantiation

LaTeX class: neustein, 2021/10/06 13:35:48 neustein driver • file: neustein09.tex (svetlana) • Prn: 2021/11/05;11:06:57 • p. 38
pro: NEUSTEIN • model: a10 • color: 4c • ref-style: Numbered (Style 1) • lang: en-US • docsubty: CHP •