Visual Syntactic Structures

• Towards a Generative Grammar of Visual Language •

By Neil Cohn

www.emaki.net **neil@emaki.net**

Table of Contents

Preface

Throughout the course of the long and traveled study of language, perhaps no work has emerged more significant in its implication for the human mind than that of Noam Chomsky's *Syntactic Structures* (1957). The crux of this work, and many following it, held that underlying the "surface structure" of linguistic utterances lay tacit "generative" principles of mental states that exemplify the linguistic competence of a language speaker (Chomsky 1957,1965). Effectively, by probing for rules lying beneath the exterior of language's sequence, its syntax, Chomsky placed linguistic inquiry as "mentalistic," concerned with discovering the tacit mental processes underlying actual behavior, rather than merely providing a descriptive account of the behavior itself. By examining the mind in this way, Chomsky's work aided in ushering about the "cognitive revolution," and changing the perception of both language and linguistics in its wake (Harris 1993). Though the theories have undergone considerable, and expected, critique and change, the issues raised by Chomsky have continued to fuel the study of linguistics as some of the most fundamental concerns in the study of language.

Indeed, as will be seen, such considerations constitute the driving force behind this work as well. While the implications and results of this present discourse cannot be speculated upon, a similar methodological undertaking is expounded herein, though applied to perhaps an unexpected area to the study of linguistics: the production of *sequences* of images emerging directly from the human mind. This *visual language* exhibits markedly different *representations* of conceptual phenomena, while still governed under the same *properties* as symbolic linguistic structures. Because this involves no visual input in order for production to take place, that being drawing an object with a referent clearly visible, the output must be a reflection of conceptual processes– and thus mentalistic. However, this can be true of individual images as well as those produced in sequence. As with any language, at the heart of investigation into visual language's linguistic theory is its syntax– the combinatorial features of language that determine which elements can or cannot properly work as a coherent sequential production.

This work then, looks to investigate these combinatorial features of visual language syntax. This will be carried out in two parts. In the first, the theory of visual language, as well as its syntax, will be put into greater context by examining its theoretical development, especially with relation to that of aural language. This will follow with a second section outlining the foundations for a working model of syntax based on the generative grammars developed in linguistics since the 1950s.

§1. Early Theories of Visual Language Syntax

§1.1 Transitional Syntax

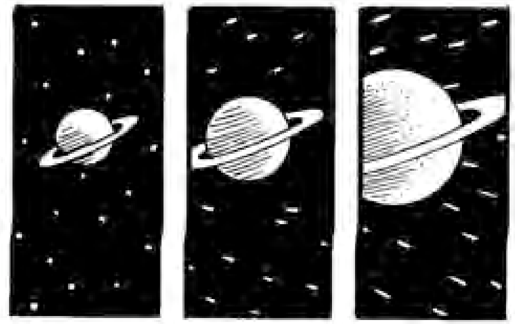
At the outset, a clarification must be made in referring to what is meant by "syntax" for a visual language. In linguistics, syntax is the field of the study that analyzes the grouping of words and phrases into sentences. In other words, it examines the structural organization of meaningful units of a language into a combinatorial sequence. In terms of visual language, this same definition is adhered to. Visual language syntax is the study of the structural organization of a sequence of images. This departs from the standard perception of examining a so-called "language of art" or semiotic studies of images, which look at syntax as the compositional makeup of a single image (Gombrich 1961, Duncan 1999). Such an approach does not adhere to a strictly linguistic meaning of the term, and is inappropriate for the study of visual language, which features purely linguistic properties, though they may be tempered to its distinct visual modality.

Perhaps it should be of little surprise, but the development of descriptive analysis for visual language grammar occurred historically similar to that of symbolic language. Like the pre-Chomskyan inquiries into syntax, visual language theorists began by looking at linear strings of syntactic units at face value.

Greatly influenced by the instructional and analytic work of Will Eisner (1985), this first venture into examining the sequential nature of visual language analytically emerged from author and theorist Scott McCloud, regarding such phenomenon in the social context of "comics."[1] While still not addressing it directly as such, McCloud did compare the form to "language" metaphorically, astutely recognizing the study of sequence as "grammar" (McCloud 1993:67), and subjecting it to a rigorous dissection which yielded a surprisingly structuralist analysis. Like the pre-Chomskyan grammarians, McCloud categorized strings of visual images into types of "state-to-state transitions," providing taxonomic distinctions for these "panel transitions," based on the shift of change between the contents within each pair of juxtaposed images (McCloud 1993). In this initial analysis, McCloud identified six types of "panel-to-panel" transitions: 1. Action-to-action, 2. Moment-to-moment, 3. Subject-to-subject, 4. Scene-to-scene, 5. Aspect-to-aspect, and 6. Non-Sequitur.

Moment-to-moment transitions featured a short interval of time passing (Quotes and examples from McCloud 1993:70-72):

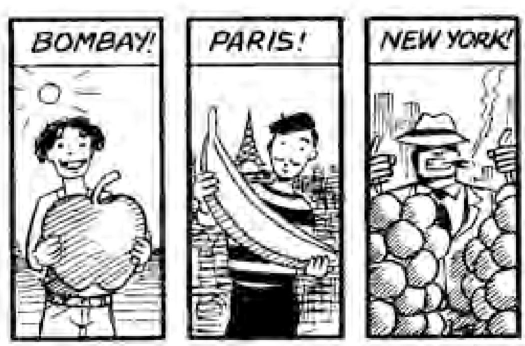
[1] The relation of comics to visual language can be delinated in a metaphorical relation to Chomsky's (1986) distinction between "E(xternal)-language" and "I(nternal)-Language." Comics (E-language) can be seen to encompass a larger sociologic and cultural construct, whereas visual language (I-language) represents the structural properties often associated with them. Thus, comics– the objects– often *contain* visual language, while, in a sense, comics as a social community and institution make up the language group that uses visual language (Cohn, forthcoming).



Similarly, Action-to-action transitions featured a "single subject" in distinct progressions of action:

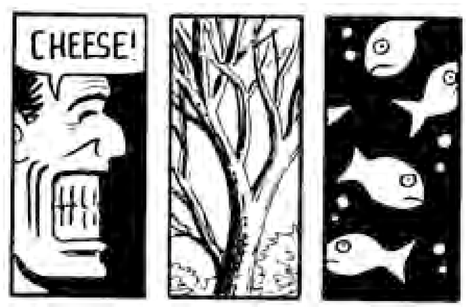
The third type of transition moved from "Subject-to-subject while staying within a scene or idea":

Scene-to-scene transitions were said to move across "significant distances of time and space":



The fifth transitional type of Aspect-to-aspect reputedly bypassed time to place a "wandering eye on different aspects of a place, idea or mood":

The final transitional type, Non-Sequitur, offered "no logical relationship between panels whatsoever":



Though he proceeded to apply his classifications to note the similarities and differences of these transitions found throughout varying works, McCloud openly admits that his categorization was an "inexact science at best" (1993:73). Indeed, his breakdown is rife with ambiguities, the quantification of time at the forefront. Additionally, since every transition type is a balanced "panel-to-panel" pairing, it would imply that McCloud is really proposing what *individual* panels should be classified as (with the exception of the temporal relation pairings), though only giving them in ways that transition between two of the same panel types. For instance, would a panel that contains an Aspect next to one that contains a Subject be a "Subject-to-subject" transition, because it moves into an *individual Subject panel*? Or, would this situation be deemed a variant "Aspect-to-Subject" transition? Despite the ambiguities found in the given categorizations and the absence of a clearly defined methodology of analysis, McCloud's contribution to visual linguistics undeniably succeeded in one extremely important respect: he provided the first comprehensive descriptive analysis of the form.

§1.2 Results and Revisions

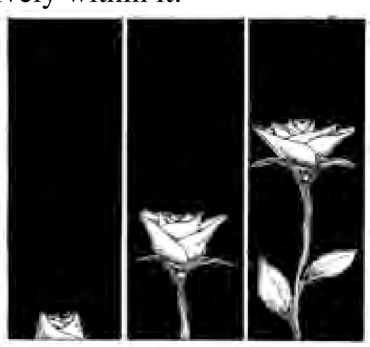
McCloud's work was far-reaching and vastly influential, and comprehensive responses took time to surface. While some praised it for the applications it could have to comics creation, communications, and other media (Harvey 1996, Horn 1998, Tong 1999 respectively), critics argued against the ambiguities found in McCloud's categories, calling for clarification and further subdivisions among types of transitions (Beaty 1998, Dean 2000 Cohn 2002, 2003). Subsequent propositions revolved around the primary distinction between narrative and non-narrative transitions, an issue that McCloud's later hypotheses would tap into (McCloud 2000a).

For instance, one subsequent breakdown addressed the time versus space ambiguity, attempting to divide these transitions, plus others into three distinct groupings:

1) Temporal transitions, 2) Spatial transitions, and 3) Spatio-temporal transitions (Dean 2000). The first category of Temporal transitions is meant to subsume relations that mark the time's passing within a held visual space, including McCloud's Action-to-action and Moment-to-moment progressions, as well as an unnamed type sequence which separate long intervals of time over a constant environmental space. The second class of Spatial transitions includes Aspect-to-aspect and Scene-to-scene transitions unaffected by the passage of time. The final grouping of Spatio-temporal transitions would include Subject-to-subject and Scene-to-scene transitions that mark a passing across both time and space.

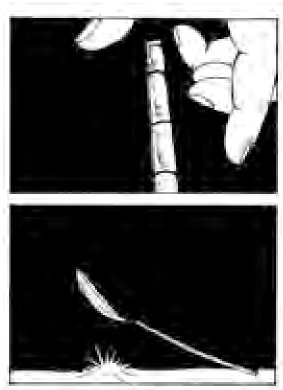
However, while these groupings attempt to avoid some of the ambiguities found in McCloud's taxonomy, they still maintain the focus of panels being momentary increments of narrative. Because of this, the approaches remain far from mentalistic, though McCloud did attempt to incorporate a mental component in the form of the pseudo-Gestalt principle of "closure" which allowed the reader to "mentally construct a continuous unified reality" out of a linear sequence of images (McCloud 1993:67).[2] Furthermore, the ambiguities relating to role of time itself are not addressed, only expanded by distributing its properties to different affected elements– either a subject or a space, which leaves the actual occurrence of time unresolved.

A reconciliation of this perspective was attempted in yet another taxonomy of transitions, clarifying the role of time, and abandoning the "panel-to-panel" binary in terminology. To do this, transitions measured the relation of one panel moving *into* another juxtaposed panel, while also identifying individual panels themselves. This classification divided into two major transitional categories with several sub-fields: 1. Temporally Progressive and 2. Temporally Ambiguous (Cohn 2002, 2003). This first category marked any panel transition displaying a progressively temporal relationship. Of this, there are two types: 1. Moment Transitions, in which time was the dominant factor allowing an action occur passively within it:



2. Action Transitions, in which an action occurred and pushed time because of it. Because occurrences must inherently work within a temporal span, this transition type exploits such a driving force of time:

[2] While the idea of "closure" attempts to give the mind at least some role in the construction of meaning from sequence, its deficiency lies in that it provides a process to explain for what *isn't* there, rather than a process to explain what *is* there. By explaining how "what is there" appears, as is the intent of this piece, the issue of explaining "what isn't there" becomes a non-issue.

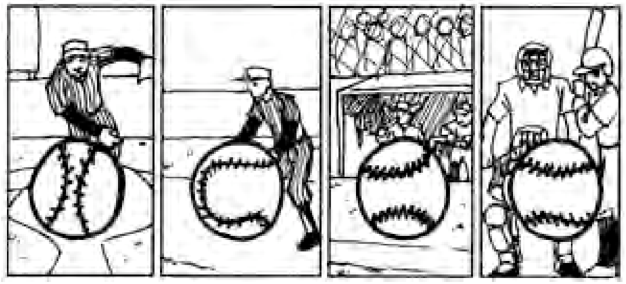


Thus, here, the explicit differences between how time is represented between panels is expressed: Moment Transitions begetting a passive account of the occurrences within the panels, letting time itself dominate, whereas Action Transitions place the focus on the occurrence itself (Cohn 2002, 2003b).

The second category of Temporally Ambiguous transitions incorporates multiple subcategories into which transitions displaying no *inherent* temporal relations are placed, though temporal constraints may apply in the context of an actual discourse. The first subcategory, Environmentally Existential transitions, based relations around the components of a single scene and environment. Into this first category fell three transitions: 1. Subject Transitions, which shifts into a panel containing a primary acting agent of the scene.

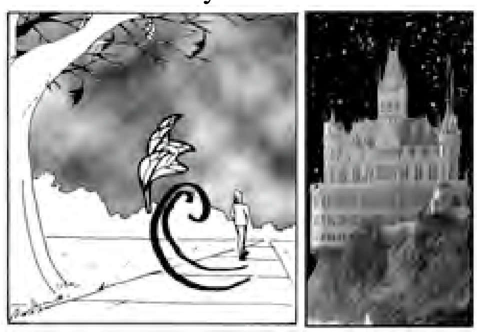
2. Aspect Transitions, which shift into a panel containing "non-acting" elements of the scene or environment.

3. View Transitions, which change the perspective at which the elements of a scene are viewed, while not changing temporally.

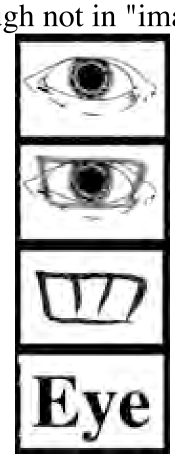


Within the Aspect and View Transitions, further divisions are made (see Cohn 2002).

The second subclass of Temporally Ambiguous transitions featured Environmentally Codependent Transitions, based on the relations of two environments and their enclosed actions, was reflected solely in a Scene Transition:



Next, followed a class of Environmentally Ambiguous Transitions, including a Cognable Transition, featuring a transition into a panel devoid of connection to an environment though retaining a semantic connection. This could apply to both visual material outside of the environment, as well as text, which contains conceptual information expressed visually, though not in "image" form:

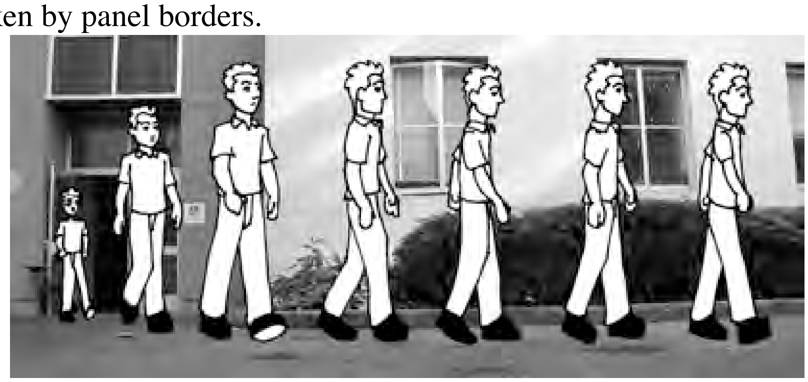


Finally, under this class, the McCloudian Non-Sequitur is maintained for seemingly unrelated transitions:

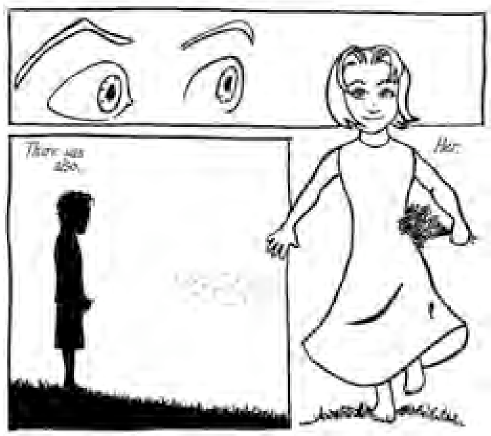
Additionally, this categorization of transitions also attempted to reconcile three previously unmentioned conditions that varied beyond the outlined transitions. These conditions were as follows:

1. Inclusionary transitions, which featured panels within panels– and thus concepts within concepts.

2. Embedded transitions, which contained transitions within the framework of a single image, unbroken by panel borders.

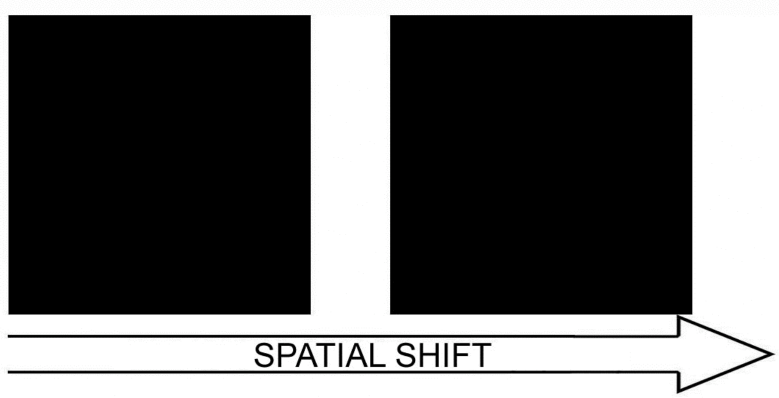


3. Overlays, where an element of the sequence intersects another in such a way that it affects multiple panels that it touches.

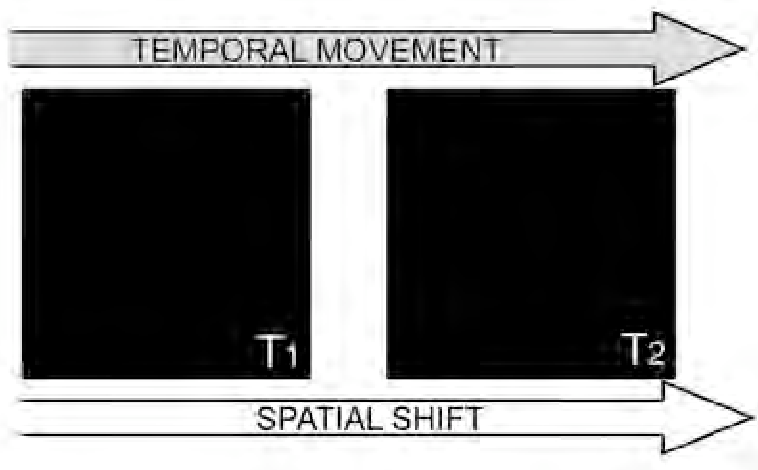


Beyond attempting to resolve the ambiguities in the early approaches, this was also the first to attempt taking into account "grammatical" roles of elements of the scene, an approach treating the panel contents as visual concepts rather than narrative increments. This stance reveals itself in the opposition of Subjects as "acting agents" versus Aspects as "non-acting elements"– which requires a sequence of panels to delineate such roles, as well as the inclusion of Cognables as non-environmental conceptual panels.

On a separate track, McCloud's next proposition of "syntax" attempted to push beyond the categorical descriptions of panel relations, and towards what he saw as the "essence" of "comics:" the Temporal Map (McCloud 2000a). Here, McCloud uses the examples of substance-free juxtaposed black squares to further the abstraction of transitional elements. The first condition assumes that these are two separate squares, and thus a move from one panel to the next would represent a shift in physical space– literally.



However, if the two juxtaposed squares represent the same square, just at two separate moments, the reading progresses in a transition across both space *and* time.



McCloud termed this equation of "space equals time" (McCloud 1995:65) as a *temporal map*, claiming it to be found not only in "comics" (re: visual language), but also in representations as diverse as sine waves of the "Orbital patterns of Jupiter's Moons," "Musical and dance notations," and diagrammatic recordings of "Sunspot activity by decade" (McCloud 2000a:207).

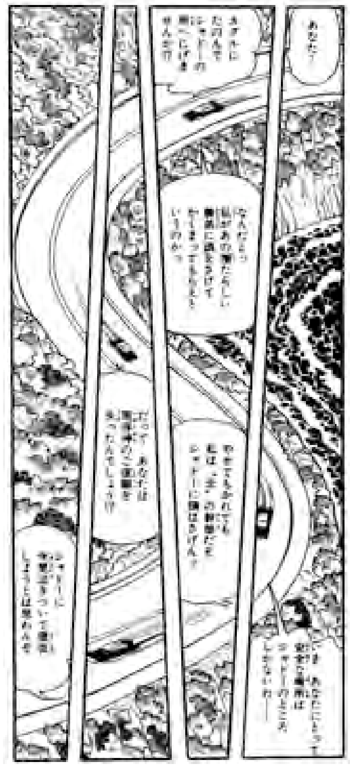
While McCloud's proposition is admirable for attempting to transcend descriptive taxonomies to a more abstract basis of analysis, ultimately, this proposal faces problems in multiple areas. First and foremost, McCloud's two conditions are held on two separate quantifiable planes. The shift in space to which he refers is a shift in *physical* space, evident in his instructions for the reader to run their finger from panel to panel to experience the spatial orientation. However, the shift in time is an *abstract* sense of time, found only in the (mental) semantic meaning of the squares' content. Truly, the "physical" sense of time would be equally applicable to both sets of conditions, regardless of the abstract temporal relations between the contents of the panels.

The second problem arises in situations that *do* meet the first condition and *are* devoid of a temporal relationship– such as any in the Temporally Ambiguous class of transitions. These obviously do not meet the temporal mapping criteria, though examples such as these will inevitably arise in production. Additionally, examples that meet the criteria of temporal mapping though blatantly defy its principles become problematic and paradoxical.

Take for example this sequence from the Japanese book *Hi no Torii [Phoenix]* (Tezuka 1986: 97) (read from right to left):

While this sequence does feature a temporally progressive relationship between the first two panels, and again with the fourth and fifth, the procession of people becomes problematic to the flow of time. Here, the interaction of the overlay between the first two panels masterfully shows the movement of the procession, by moving the land under which they are passing. Essentially, when reading the first panel, the figures in the second panel are behind those of the first with regard to that space. However, by the time of the second panel, they now occupy the space of the first panel. These first two panels pose no problem to the temporal mapping theory, and, in fact, execute it with stunning brilliance. However, the fifth panel brings the problematic element to the fore, because the people in the fifth panel should be entering the same space as the first two panels in order to traverse across time and space. This is not true though, since those people are walking through a *space* that *precedes* the second panel's, while occupying a *time* that occurs after the second panel's. Essentially, if the panel increments were treated strictly as moments this interaction creates a temporal paradox.

A different type of problem arises again from another volume of Tezuka's series (Tezuka 1987b: 308) (read right to left):



In this example, the movement of the car weaves in and out of each panel, creating individual panels that contain two separate moments of the narrative– moving both forward and against the flow of the panels as it slaloms along the path of the road. If this example was truly an instance of temporal mapping, time would be moving forward, then jump backward, then forward again through the panels– which obviously does not occur, as indicated by the constant forward motion of the car on the road and the progression of the dialogue. Once again, this sequence would create a paradox to a temporal mapping analysis, one that could have been simply avoided if the divisions between panels were made descending horizontally rather than vertically.

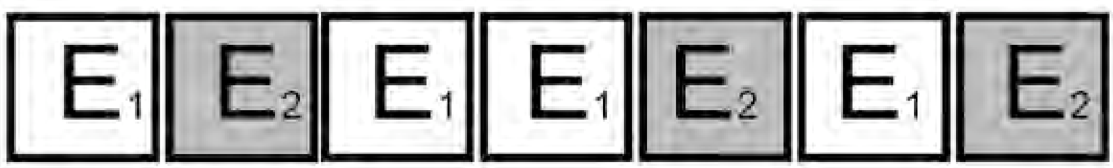
If the contents of panels are indeed treated as momentary increments, how is it that these temporal paradoxes can still be read and understood with very little mental strain? Once again, like earlier taxonomies of panel transitions, temporal mapping finds its deficiencies in treating panels as temporal elements, rather than as conceptual units (Cohn 2003b, 2003).

Since it arises throughout the discussion, a note about the mentalistic perspective of temporal elements should me made. This conceptual view of these increments does not perceive such elements as actual momentary increments of time. Rather, they are perceived as representations of conceptual information conveyed visually, in which causation or predication is manifest appearing similar to time’s passage because of the subject matter being represented. However, because this shift is abstract and cognitive, no actual temporal designation should be implied as inherent to the system itself, despite the temporally bound vocabulary that describes them for the purpose of jargon transparency (Cohn 2003a).

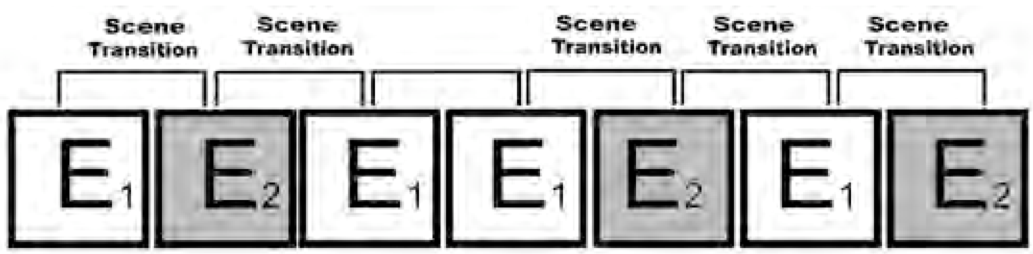
While examining the problems found in the "temporal paradox" examples, in light of moving towards a conceptual perspective, the issue of how analysis is made also arises. Though he describes measuring transitions by his original categories of panel transitions, McCloud never elaborates on a methodology (McCloud 1993). In establishing such a method, several issues provided challenges to the recording of a linear approach to syntax, most notably: 1) How are transitionally variant conditions (such as Overlays) measured? and 2) How does one account for panels that connect to more than one other panel while retaining a linear reading?

In order to overcome these challenges and establish a workable method of analysis, a formal system of noting panel transitions was arrived at by Cohn (2003a) in the form of *web diagrams*, based on his transitional taxonomy. As will be evident, though the web diagram method does provide a reasonable answer to these concerns to some degree, like symbolic language, it is the lattermost dilemma that would eventually signal the downfall of visual language's transitional syntax.

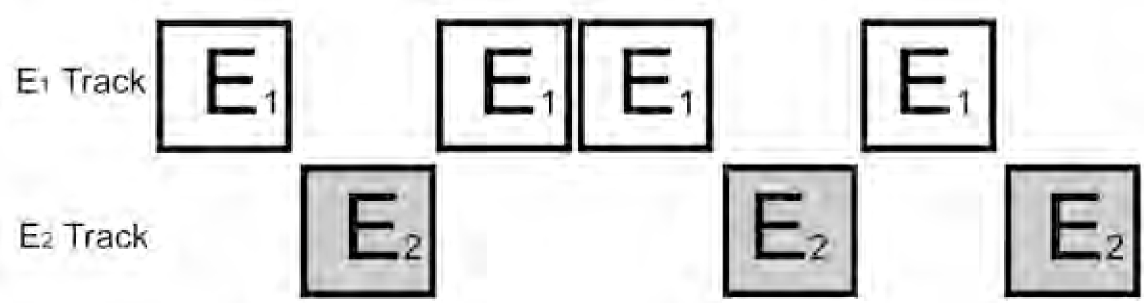
The concern of panels with multiple connections first arose in the context of instances of "parallel cutting," scenes in which two or more environments occur simultaneously in a single linear stream.



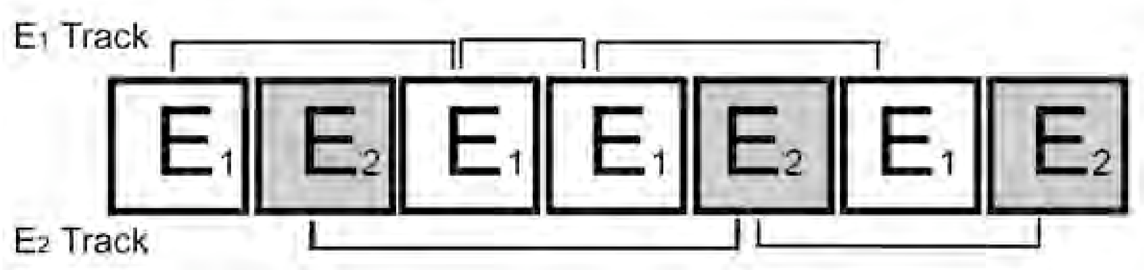
Here, two environments (denoted by E_1 and E_2, with E_2 greyed for further clarification) are intermixed and read in a single linear strand, where two temporal situations are assumably occurring at the same time. If analyzed linearly, every time that a transition occurred between the scene in E_1 and E_2, by definition, a Scene Transition would occur:



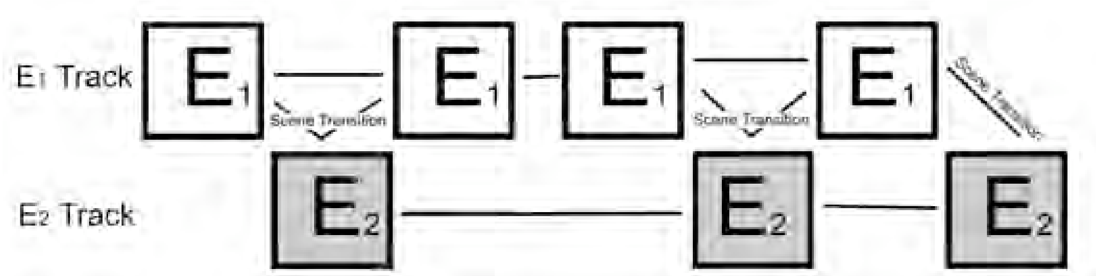
However, this posed a problem, since mentally, one retains knowledge of the scenes as they unfold, and understands the connections made the separate environments' occurrences. Essentially, here the two environment's scenes become split into two separate paths.



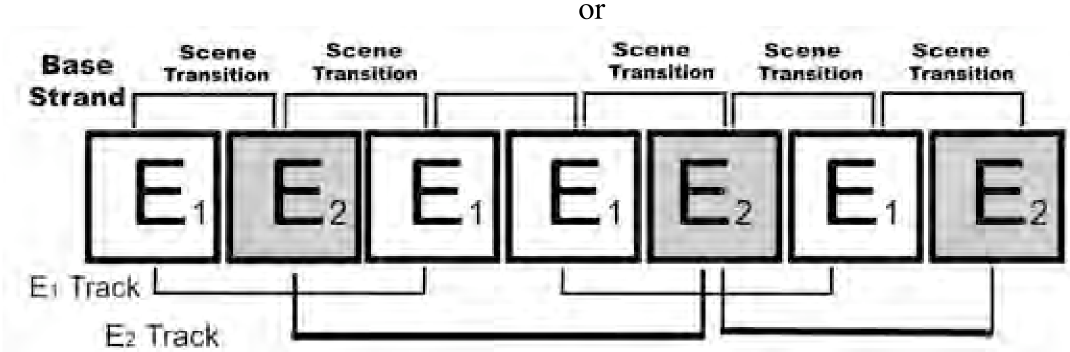
This dual mental track allowed for an enclosed environment's occurrences to be maintained and connected.



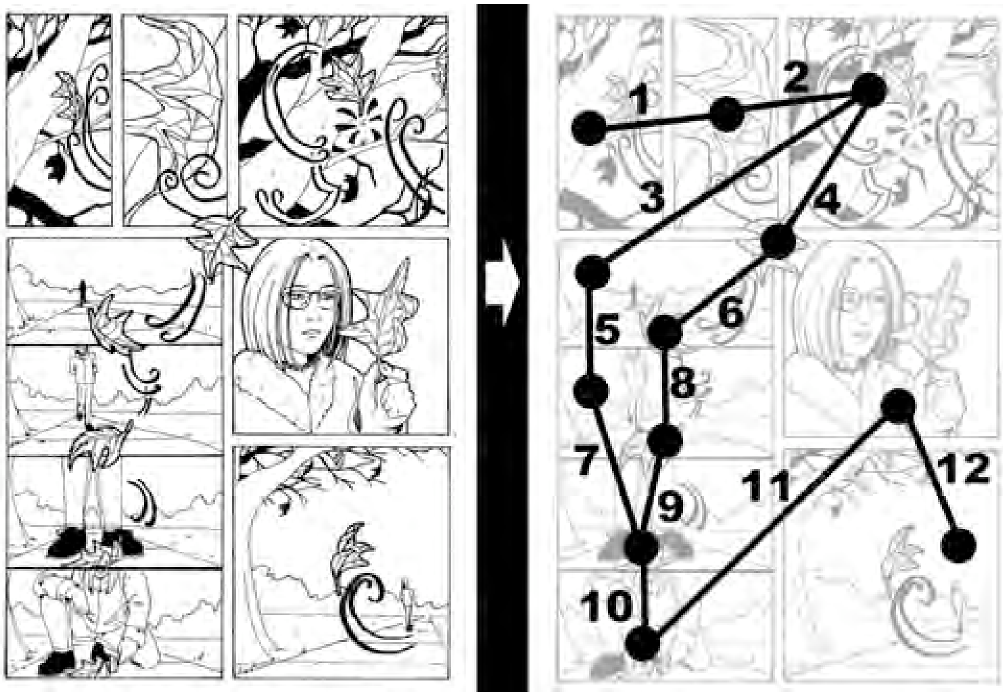
However, the juxtaposition of the panels in their original linear strand did not go overlooked, as one is aware that the shift in scenes transpires, despite its secondary nature to the primary events within each of the environments. Thus, for each panel that is involved in this Concurrent Streaming Environment (CSE), two transitions would be recorded and analyzed.



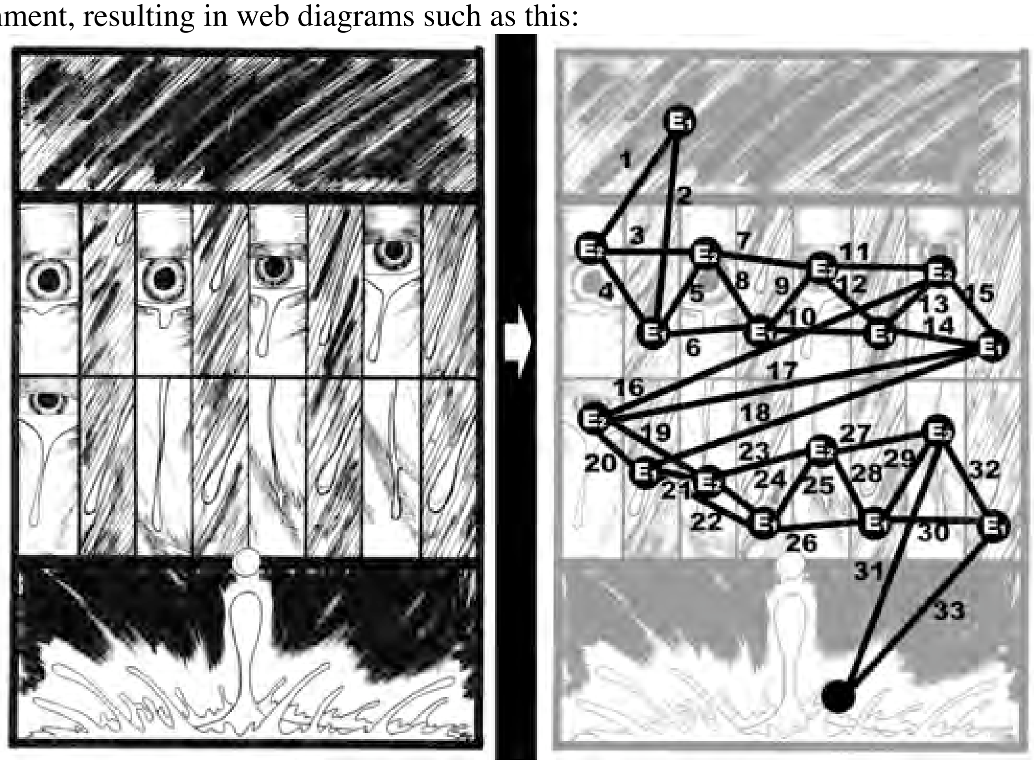
or



CSE transitions could then fall under the larger group of Multi-Engaged Transitions (MET), which could address the concerns of the variant conditions of Inclusionary, Embedded, and Overlay transitions– all of which could be represented by web diagrams. At each transitioning element (i.e. a panel or embedded/overlay object), a point is placed. Then, the points were connected for each transition that occurs, which can be counted and recorded. For a linear strand, a web diagram will be very simple to create and understand, though more complex transitions beget increasingly complicated analyses. For example, in this scene involving an Overlay, the web diagram would be as follows:

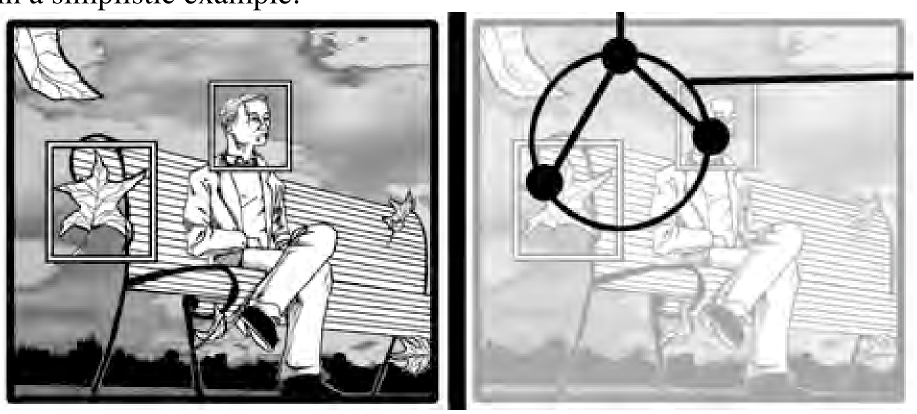


Similarly, a CSE notation would include markers within the dots to indicate Environment, resulting in web diagrams such as this:

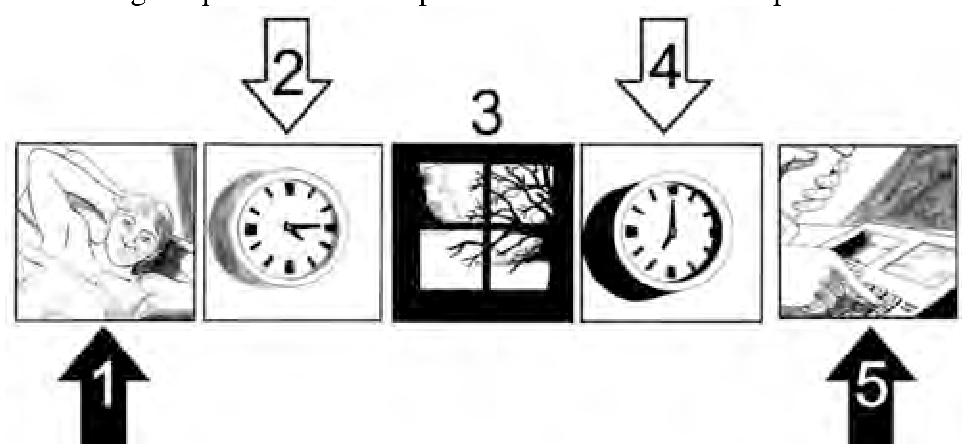


Further development of the web diagram analysis included methods to account for "Divisional" forms of Aspect Transitions, which dealt with issues of image constancy. This introduced the component of a "Unifier" (represented by a circle), and could address situations of varying complexity, such as here:

The Unifier also helped to account for Inclusionary transitions as well, here shown in a simplistic example:

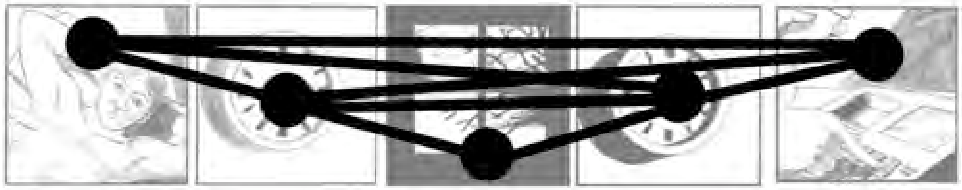


Though the web diagram method seemed to resolve issues involving nearby transitioning elements, situations with multiple connections requiring greater distance and additive knowledge require a more complex model. Take this example for instance:



Here, nearly every panel would have to be considered at MET panel. Most glaringly, the second panel would have to connect to the fourth panel, because of the overt change in time between them, considered an Action Transition. Next, the first and last panels would need to be connected because of the obvious separation in time between the two of them as well. Further complication arises if one questions the relations of the additive knowledge of the first four panels. Despite the ambiguity in time between the first and second panels, because the first and fourth panels are separated by a panel denoting time (the second), it must logically be deduced that the first and fourth panels could be connected by some increment of time as well.

The resulting web diagram would look something like this:



The problem with this should be evident. Though the first panel must logically *connect* to the fourth (and fifth), can it really be said to engage in a *transition* with it? Furthermore, because no temporal change is indicated between certain panels (say, any panels between the second and the fourth) doesn't that mean that an infinite amount of non-temporal marking panels could potentially be put in between them? And if so, would the "transition" between those panels (two and four) still exist if say, twenty non-temporal panels separated them? Clearly, there must be an easier method of analysis that account for this additive knowledge, or else every panel could very well become connected to every other panel, all the time– which by a transitional sense would be far too difficult and cumbersome, if not false in meaning of "transition."[3]

§1.3 An Alternative Perspective in a Traditional Field

This situation can be given further clarity by examining the similar critique made by Chomsky of the transitional styled grammar of aural language. Prior to Chomsky's innovations, syntax was looked at as a state-to-state grammar (Chomsky 1957, Harris 1993), whereby a word of one type governs the selection of what type of word follows it. For instance, it is likely that a noun (ex. woman, iguana) or an adjective (ex. big, tired) will follow an article (ex. the, an), but not a verb (ex. sang, ate). Thus, a certain probability would exist for each transition between word types.

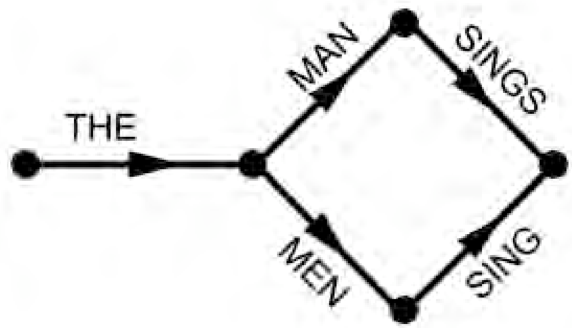
By this account, as a person chose a word, it would contextually dictate which word would follow by transitional probability, the result of which was known as a "first-order Markov chain" (Corballis 1991). However, this type of generation does not necessarily produce acceptable sequences, as the example shows:

> There are green Frisbees sail along joyfully crossing the chicken feathers on houses of everyone…

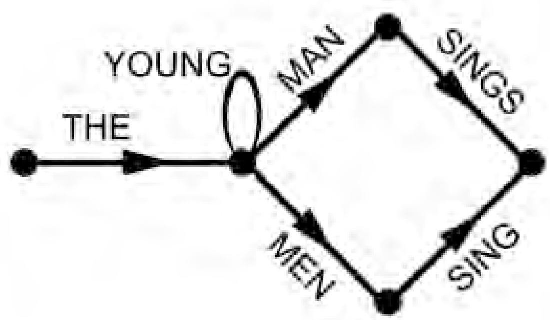
The sentence as a whole is essentially nonsense, though each pairing of words is acceptable. To remedy this, it was thought that by increasing context, more acceptable sequences would result. Thus, rather than pairings relying simply upon the immediately preceding word, a group of four words might provide the context for the fifth, as in a "forth order Markov chain." This type of "finite-state grammar" adequately expressed grammatical constructions up to a certain point, and was expressed through a "state

[3] This is not to mention other problems with a transitional approach, such as transitional ambiguity. For example: Does a transition showing a change in time, but moving to another subject register as a Subject Transition or an Action Transition?

diagram" (Chomsky 1957), such as this one for the alternate sentences "the man sings" and "the men sing":



By tracing the arrows moving from the point on the left to the right, the probabilistic states of the sentence could be followed. Hypothetically, such a method could be applied to the previous web diagrams as well, and the similarities in systems should be self-evident, though the former played no role in influencing the creation of the latter. Additionally, if a variable component were to be added between parts, a "closed loop" would be added, such as this one to create "the young man sings" or "the young men sing":



Given this context, we can understand that the transitional syntax proposed by McCloud (1993, 2000a) is essentially a "first order Markov chain" model, with subsequent taxonomies incorporating web diagrams (Cohn 2003a) progressing to account for higher ordered chains. As was implied by the visual language analysis, this method allows for grammatical sequences to occur only up until a certain point. Observe a sentence such as this:

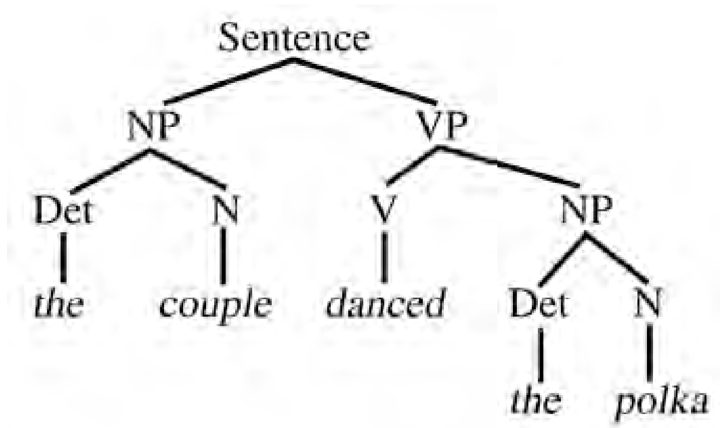
> My *roommate* who recently fell rather ill and needed to stay home from class today despite his midterm *enjoyed* watching TV.

Here, the italicized words "roommate" and "enjoy" are clearly linked. However, they are separated by sixteen other words, thus requiring a "seventeenth order Markov chain." Such a grammar would render unwieldy performance and processing problems for any language user, not to mention for a language learner (Chomsky 1959). Furthermore, natural language users can produce sentences of infinite length, a feat incapable by definition of a finite-state grammar.

To remedy this problem, in his book *Syntactic Structures* (1957), Chomsky proposed a set of computational rules that would allow for this “long-range connection” and sentences of infinite length. Observe the following “phrase structure” rules:

(i) Sentence→ Noun Phrase + Verb Phrase
(ii) NP→ Determiner + Noun
(iii) VP→ Verb + NP

One can read this, as in (i), as “a Sentence *consists of* a Noun Phrase *and* a Verb Phrase.” Chomsky termed this a "generative grammar," referring to the way in which a system of rules can *generate* an infinite array of acceptable sentences (Chomsky 1965, Harris 1993). Generally, these structural descriptions are represented by a *tree diagram*, or labeled brackets:



$[_S [_{NP} [_{Det}\ \text{The}]\ [_N\ \text{couple}]\ [_{VP} [_V\ \text{danced}]\ [_{NP} [_{Det}\ \text{the}]\ [_N\ \text{polka}]]]]$

Beyond these generative procedures, Chomsky proposed an additional component to his grammar: transformations. This component allowed for structural change of a sentence. Consider the following sentence:

What are you doing?

Because the phrase structure rules cannot fully explain such a construction straightforwardly, a transformation would alter the string to be capable of generating such a sentence. Thus, a sequence would have two syntactic structures, its transformed surface form, and an underlying form:

Surface form: What are you doing?
Underlying form: You are doing what?
Structural change: NP_1-aux-V-NP_2→ NP_2-aux-NP_1-V

In this case, the transformation moves both “are” and “what” to the beginning of the sentence, thereby giving the grammar considerable power for generating acceptable sentences. The result of such an analysis placed these generative principles tacitly within the mind, putting linguistic theory in the stead of discovering these mental processes, rather than merely providing descriptive accounts of the external behavior that results from them (Chomsky 1965, Harris 1993).

Just as this approach works for symbolic aural language, it does for visual as well. Granted, since its inception in the fifties and sixties, generative grammar has undergone extensive changes, both with regards to generative procedures and the transformational component, not to mention additional rules and constraints to account for a variety of phenomena that occur in language. However, since the task herein is to move from a state-to-state to a generative grammar for visual language syntax, we will start at this familiar point of origin.