Crosspower: Bridging Graphics and Linguistics

Haijun Xia
University of California, San Diego
haijunxia@ucsd.edu

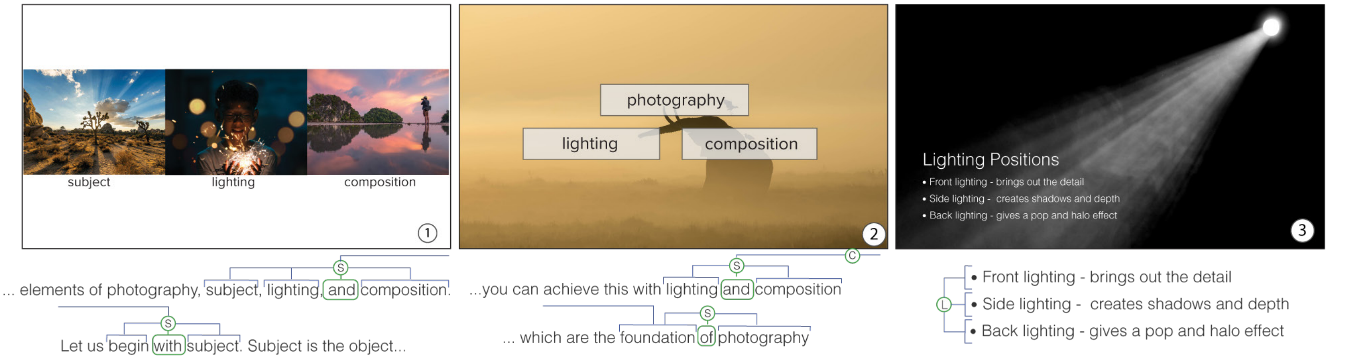

subject
lighting
composition
1
photography
lighting
composition
2
Lighting Positions
• Front lighting - brings out the detail
• Side lighting - creates shadows and depth
• Back lighting - gives a pop and halo effect
3
S
... elements of photography, subject, lighting, and composition.
S
Let us begin with subject. Subject is the object...
C
S
...you can achieve this with lighting and composition
S
... which are the foundation of photography
L
• Front lighting - brings out the detail
• Side lighting - creates shadows and depth
• Back lighting - gives a pop and halo effect

Figure 1. With Crosspower, a user can directly interact with the linguistic and organizational structures in a script or outline and use them to create graphic elements and compose graphic effects. (1) graphic layout indicated by syntactic conjunction structure, (2) layout indicated by the "foundation" semantic structure, (3) graphic list indicated by list structure in the script.

ABSTRACT

Despite the ubiquity of direct manipulation techniques available in computer-aided design applications, creating digital content remains a tedious and indirect task. This is because applications require users to perform numerous low-level editing operations rather than allowing them to directly indicate high-level design goals. Yet, the creation of graphic content, such as videos, animations, and presentations often begins with a description of design goals in natural language, such as screenplays, scripts, outlines. Therefore, there is an opportunity for language-oriented authoring, i.e., leveraging the information found in the structure of a language to facilitate the creation of graphic content. We present a systematic exploration of the identification, graphic description, and interaction with various linguistic structures to assist in the creation of visual content. The prototype system, Crosspower, and its proposed interaction techniques, enables content creators to indicate and customize their desired visual content in a flexible and direct manner.

Author Keywords

Language-Oriented Authoring; Text-based Editing; Natural Language Processing; Reification

UIST '20, October 20–23, 2020, Virtual Event, USA © 2020 Copyright is held by the owner/author(s). ACM ISBN 978-1-4503-7514-6/20/10. https://doi.org/10.1145/3379337.3415845

INTRODUCTION

Today, graphic content such as videos, animations, and presentations can be produced at increasingly fast speeds thanks to the proliferation of computer-aided design applications [17, 48]. The user interfaces found within these applications follow the principles of direct manipulation, allowing content creators to directly manipulate and configure graphic elements [44].

Despite the support for such directness, creating and editing graphic content remains a laborious process because content creators must manually configure a multitude of properties to achieve even the simplest graphic layout or animation. Imagine if you were a photographer making a tutorial video and this was the first sentence of your script, "*Today, I will talk about three key elements of photography, subject, lighting, and composition.... let's begin with subject*" (Figure 1-1). To design animations to accompany this script, you may want to have images representing subject, lighting, and composition appear one by one and then the image of the main element, *subject,* to be highlighted. To do this, you would need to (1) manually find the images of the mentioned elements online, (2) copy these images into a canvas, (3) resize and crop each image to the same dimensions, (4) align the images on the canvas, (5) create an 'appear' animation for each image, and (6) create a 'highlight' animation for the subject image.

Creating this first segment of your video is tedious because there are many repetitive tasks and image properties that need to be manipulated. However, this segment uses images and graphic effects to enhance engagement and facilitate comprehension within viewers [1, 50]. In particular, you have taken specific care to ensure the layout, ordering, and animations in the segment correspond to the content in the script, i.e., the images of the elements that are resized and aligned on screen are direct mappings of the "*subjects, lighting, and composition*" conjunction phrase that is in the script and the order in which the images "appear" is informed by the order that the element occurs within the phrase. Similarly, the highlight animation corresponds to the phrase "*begin with*", whose semantics suggests a transitional action and provides direction about the order in which you wish content to change.

The above correspondences between the visual artefacts on the screen and the script that you had prepared is an example of the Congruence Principle, which has been recommended for effective visual communication, "*the content and format of the graphic should correspond to the content and format of the concepts to be conveyed*" [50]. It also demonstrates how the syntax and semantics found within language can be used for language-oriented authoring to inform graphic content that will be created. Imagine if content creation system could automatically provide templates informed by the syntax and semantics within a script, where images are automatically resized and aligned, and transition animations to be added whenever "*begin with*" or similar phrases are encountered. Such abstracted and encapsulated functionality would allow users to directly indicate their high-level design goals once, in a form that is more natural to them than manually performing a series of tedious low-level editing operations.

While research has been exploring the use of natural language input to create graphic content (e.g., 3D scenes) [8, 11, 27] their main theme has been the literal conversion of highly descriptive, domain specific language. Further, the pursuit of fully automated processes inherently makes the linguistic elements of the language inaccessible for further customization or editing. The central theme of this project is thus to explore language-oriented authoring, i.e., leveraging the latent structures inherent in language to facilitate the creation and manipulation of graphic content.

This work contributes a systematic exploration towards this goal. First, we explored the linguistic and organizational structures that can be extracted from written content with Natural Language Processing (NLP). Second, we defined a language-driven grammar that describes these structures using visual layouts and animations. Third, we designed interaction techniques that enable content creators to access and leverage these structures while creating graphic content.

To explore the utility of language-oriented authoring, a research prototype, Crosspower, was developed. Crosspower supports users in quickly navigating, selecting, modifying, and combining the structures in language to compose and adjust graphic layouts and animations. Interaction techniques that were designed to complement Crosspower can significantly reduce manual effort during graphic content creation, while enabling rapid and flexible customization. The new paradigm explored in Crosspower can be broadly applied to videos, animations, presentations, and other media, as the production of such content, despite their visual basis, often begins in a written form (i.e., as an screenplay, script, or outline). Written language allows visual content creators to communicate ideas, concepts, and stories, but also plan, prepare, and prototype visual forms, with minimal costs. Crosspower leverages the precedent role of language in existing creation processes and unleashes new power with language. The results of an expert evaluation of Crosspower demonstrated that the use of language structures with the proposed interaction techniques enabled users to easily indicate and customize visual content.

RELATED WORK

This research draws on prior work on visual content generation from natural language, natural language user interfaces, language-based authoring and editing, mapping between visual content and language, and extending direct manipulation through reification.

Visual Content Generation from Natural Language

Research into the generation of virtual content from natural language has been of interest for many decades. In early research, the SHRDLU system allowed users to instruct a computer to manipulate 3D objects using natural language input [53]. Situated in a simple constrained environment, SHRDLU could understand English using a pre-defined parser to handle basic language units such as clauses, noun groups, and prepositional groups for a small set of words.

More recent research has explored approaches to automatically convert descriptive and domain-specific text into representative visual content. WordsEye, for example, converted input text into 3D scenes by matching word semantics to pre-defined functional and spatial properties of 3D models [11]. Chang, Savva, and Manning created a system that learned the relative spatial relationships between 3D models from common spatial words and phrases in text annotations using machine learning (e.g., left, right, on top of, etc.) and applied these relationships to convert new text into 3D scenes [8]. Similar machine learning techniques have also been used to generate other types of visual content from text, including 3D shapes [27], images [18], short video clips [35], and infographics [12].

The use of knowledge structures, linguistic information, and heuristic-based design process for visual content generation has also been explored. Videolization, for example, utilized a knowledge graph to automatically generate videos from Wikipedia articles [26]. Word concreteness, a psycholinguistics property measuring how closely a word is related to perceptible concepts, has been utilized to automatically convert text into slideshows by composing images of the most concrete words [33]. Crosscast utilized heuristic-based algorithms to extract relevant information from audio transcripts for travel podcasts, compose search queries, and retrieve relevant visual content to augment audio travel podcast [56].

One limitation of using automatically generated content is that the visual styles of content are limited. This is because computational models are designed or trained for specific domains. The quality of the generated content also decreases if the domain and style of the input text do not match the original model. Moreover, the end-to-end conversion from text to visual content inherently prevents users from directly accessing and manipulating the structures in the text to customize the visual outcome. Therefore, the focus of Crosspower has been to extract and expose the general organizational and linguistic structures in text and enable users to directly select the elements and structures they wish to visualize, allowing them to flexibly compose desired graphic styles.

Natural Language Interfaces

Since the pioneering Put That There system [2], the HCI community has continually strived to leverage humans' rich speech and language skills to interact with computers. The use of voice and language input has become increasingly popular in recent years thanks to improvements in speech recognition and natural language understanding, and especially so when voice input is combined with other input modalities such as mice, gestures, or pen input [2, 30, 38].

Another benefit of natural language interfaces is that they allow users to directly articulate their intended operations without learning and navigating complex user interface menus. For example, VoiceCuts enabled users to issue short, partial, and shortcut-like commands to quickly perform intended actions in creativity-based applications [30]. PixelTone enabled users to express desired operations with natural language in an image editing application [32]. DataTone [18] and Orko [46] also enabled users to issue natural language commands to explore data visualizations.

A common approach within such systems is to extract commands and parameters from natural language input and then execute the corresponding functionality. Crosspower's use of natural language, however, is not command-centric; it directly suggests graphic outcomes based on information found in natural language, reducing the number of manual operations required to achieve a given visual outcome.

Language-based Authoring and Editing

Extensive research has leveraged the shared linear temporal properties among, language, audio, and video to assist with the editing of media clips [41, 42, 45]. For example, time-aligned interactive transcripts enable audio producers to directly modify a text transcript, resulting in corresponding edits in an audio waveform [41, 42]. Quickcut utilized a time-aligned transcript of a voiceover and a transcript of annotations with raw footage to enable editors to quickly match narration story events with segments in raw footage [49]. Other systems have utilized time-aligned scripts to insert cuts and create transitions in interview videos [44], edit the speech content in talking-head videos [16], insert B-roll into main footage via interactive transcripts [22], and assist with audio recording and editing [45].

In addition to matching the linear temporal properties that are intrinsic to text, audio, and video, the focus of Crosspower has been to leverage the format and structures between graphics and semantics to assist in the bidirectional editing of text and visual content. Crosspower also shares in the spirit of using programming languages to create graphic content and behavior. Users need to conform to strict syntax rules with programming languages, whereas Crosspower allows users to "program" with linguistic structures using lightweight interactions.

Bridging Graphics and Language

Bridging the unequal expressive powers of graphics and language requires that there is a mapping between the graphics and linguistics. Such mappings are often found in databases. ImageNet, for example, was an image database that organized images using nouns in the WordNet lexical database [13, 37]. The mapping between the nouns and images resulted in a de facto visual dictionary that was used to teach computers to recognize common physical objects. The Visual Genome dataset extended this idea by using crowdsourcing to map the relationships between objects in static images to image annotations [31]. While these datasets aimed to improve our understanding of real-world photos and videos for computer vision purposes, Crosspower's focus has been on the graphic representations of the semantic relationships.

Another related project is Chalktalk, a digital presentation and communication system that enabled users to invoke animated graphic elements with gestures [39]. It used a set of mappings between pre-defined gestures and animations to enable users to compose animated elements for storytelling. Chalktalks' reliance on gestures, however, required users to learn a large gesture vocabulary. With Crosspower, the power of language is unleased using users' acquired vocabularies through a set of simple and flexible interaction techniques.

Extending Direct Manipulation Through Reification

Fundamentally, this work sits alongside the HCI community's efforts to extend the principles of direct manipulation [4, 44]. While most existing graphic user interfaces allow users to directly manipulate interface widgets using mice, touch, styli, and other input modalities, directly manipulatable interfaces are not *direct enough* if they do not support high-level design goals [4] and users must constantly perform numerous low-level operations despite with direct manipulation.

Recent research have demonstrated the benefits of reifying the various elements among users' workflows within an interface that were not previously manipulatable [4], such as attributes [55], selections [54], visual encodings [57], the spatial arrangements of graphic objects [15], or the motion trajectories of objects in videos [14]. This work reifies the various latent structures in language and enables users to flexibly and directly articulate high-level design goals by interacting with new interface elements.

CROSSPOWER

Leveraging the structures inherent in language has the potential to significantly reduce the manual effort required to create effective and congruent visual content. To explore this notion, the following three steps were undertaken:

1) ***Identification of linguistic structures.*** To leverage the latent structures in language, we first identified the linguistic structures that can inform the creation of meaningful graphic templates and operations, but also can be extracted from text using state-of-art NLP techniques.

2) ***Specification of linguistic-graphic mappings.*** To create graphic content via linguistic structures, we then needed to specify how the various linguistic components can inform the creation of graphic content and intended effects. We developed a language-driven grammar that specifies the graphic representations of a linguistic structure.

3) ***Interaction with linguistic and graphic structures.*** We then implemented a set of novel interaction techniques that enable users to directly interact with language structures and their graphic correspondences to quickly and flexibly create desired graphic effects.

The creation of a comprehensive database of mappings between linguistic structures and graphic effects, i.e., a de facto visual dictionary, can enable a wide range of applications, however, constructing such a large-scale database requires significant costs [13]. Therefore, as a technology probe and proof of concept to explore various interaction mechanisms [23], Crosspower utilizes a small pre-defined database of 152 linguistic structure templates. Crosspower was implemented as a web-application with mouse and keyboard input. Crosspower utilizes three NLP Toolkits to extract various linguistic structures due to the differences in availability, stability, and performance of different language parsing modules. In particular, the Google NLP [19], CogCompNLP [28], and Stanford NLP [34] toolkits were used to extract syntactic, semantic, and coreference structures. A time-aligned script was acquired using the forced alignment approach [42].

IDENTIFICATION OF STRUCTURES IN LANGUAGE

A language is a structured communication system that follows a grammar or set of combinatory rules to convey intent and meaning. The most basic elements of any language are morphemes (e.g., *dog, eat, -s, -ing*). When combined, morphemes form words (e.g., *dogs, eating*), which can then be further combined into phrases, clauses, and sentences (e.g., *The dogs were eating*), and then discourse. The syntax of a language dictates the allowable order in which words can be combined into sentences. The semantics of a language, however, describes the meaning or interpretation of words, phrases, and sentences, and discourse. Visual organizational structures, such as paragraphs and sections, are used in written language to visually organize semantics.

The goal of Crosspower is to leverage the structures in language that can indicate high-level graphic relationships to ease the creation of graphic content. Crosspower focuses on three linguistic structures, i.e., syntactic, semantic, and coreference structures, as well as commonly used organizational structures such as sections and lists.

Syntactic Structures

Syntactic structures, or grammars, are the low-level linguistic rules that govern the combination of words within a sentence, without giving reference to their meaning, e.g., how adjectives can describe nouns or adverbs can describe verbs. One common syntactic structure used in NLP is the dependency structure, which describes the syntactic relationship between words using binary asymmetric relations, wherein every word is associated with one dependee [20] (Figure 2-1). As an example, the word "*key*" depends on word "element" through an "*amod"* (adjective modifier) relationship.

Within Crosspower, the descriptive relationships indicated by the syntactic structures can describe the properties and relationships among the corresponding graphic elements. Through the syntactic structure, Crosspower can also extract the conjunction structure by extracting elements that are connected through a "conj" (conjunction) relationship.

Semantic Structures

Semantic structures describe the relationships between words by analyzing their meaning. While semantics are easy for humans to understand, determining the semantics of a simple sentence is a challenging task for an algorithm. Given the large vocabulary and infinite combinations of words that can be created using the English language, a common computational approach to extract meaning from a

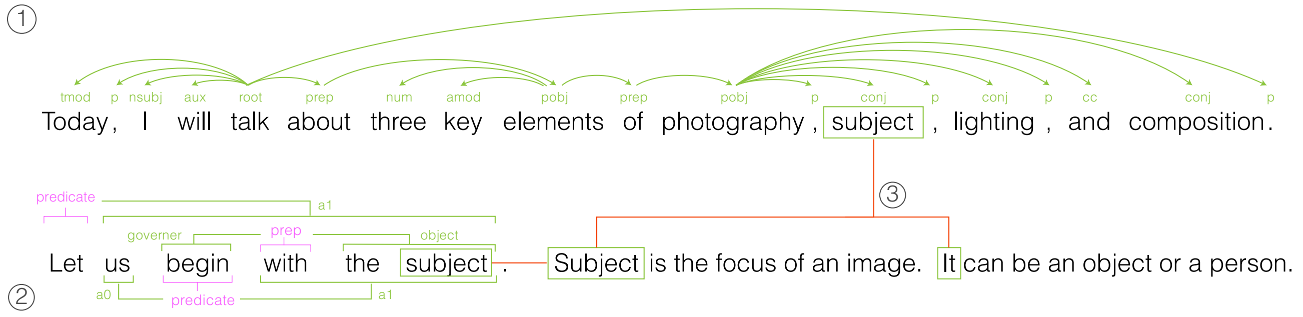

①
tmod
p
nsubj
aux
root
prep
num
amod
pobj
prep
pobj
p
conj
p
conj
p
cc
conj
p
Today , I will talk about three key elements of photography , subject , lighting , and composition .
③
predicate
a1
governer
prep
object
Let us begin with the subject . Subject is the focus of an image. It can be an object or a person.
②
a0
predicate
a1

Figure 2. The (1) syntactic, (2) semantic, and (3) coreference structures found in the example script about photography.

phrase or sentence is to examine the semantic roles of the linguistic elements in the sentence. In the sentence, "*Let us begin with the subject*", a semantic relation between *us*, *begin*, and *with the subject* can be extracted (i.e., *begin(us, with the subject)*) such that *begin* is the action or verb, and *us* and *with the subject* are semantic arguments of the action (Figure 2-2). The roles of the semantic arguments are also indicated with "a0" assigned to arguments that are agents or causers of the action, and "arg1" assigned to the patient or receiver of the action [2].

Semantic structures can not only be indicated by verbs, but also nouns and prepositions [36, 40, 46]. The variety of semantic structures that are possible within a sentence can also result in hierarchical structures where the argument of one semantic structure can contain other semantic structures (Figure 2-2). The process of determining semantic roles is called Semantic Role Labelling and can be reliably performed using existing NLP toolkits [28].

Semantic structures are used in Crosspower, in that they indicate semantic relationships that are often represented graphicly in content such as videos, animations, and presentations. By providing the graphic counterparts for the constituents of a semantic structure, and organizing them based on the semantics, Crosspower can enable users to easily create desired layouts or animations.

Coreference Structures

Understanding the flow of semantics across sentences requires the identification of coreference, which occur when multiple expressions in language refer to the same entity, either by explicitly using pronouns or by implicitly being inferred based on the context. Figure 2 shows an example in which the multiple mentions of *subject* and *it* refer to the same entity. Coreference resolution is the task of identifying words or phrases that refer to the same entity, which can be performed by NLP toolkits [34].

Within the context of Crosspower, coreference structures can indicate whether the transformations and animations corresponding to the semantics can be referred to using the same graphic elements, thereby enabling users to quickly create a sequence of graphic effects.

Organizational Structures

In addition to the structures that are implicitly embedded among the order and semantics of words, writing makes use of explicit organizational structures and rule sets to convey intent and meaning. The use of paragraphs, sections, and headings, for example, allow writing to be organized thematically, enable argumentation, enhance connectivity and flow, and provide clear visual organizations. Phrases and sentences can also be organized into lists to convey the sequential or parallel relationship amongst list items.

Within Crosspower, the extraction and utilization of such organizational structures is used alongside linguistic structures to ensure that users can consistently interact with the various structures in language.

DEVELOPING A LANGUAGE-DRIVEN GRAMMAR

Crosspower utilizes an explicit language-driven grammar to specify the corresponding graphic representations of linguistic structures. Among the three linguistic structures, syntactic and semantic structures can suggest graphic components as well as their appearance, layout, and animation relationships among various linguistic elements. For each constituent of a syntactic and semantic structure, the grammar specifies the content and form of the corresponding graphic element, with the semantics encoded in the appearance, spatial arrangements, and behaviors of the graphic elements. Coreference structures allow a user to specify whether different semantic arguments refer to the same graphic element.

Specification of Content

A semantic structure may indicate three possible operations on graphical elements based on the semantics and context: 1) the need for new graphic elements, 2) the transformation of existing elements, or 3) the removal of existing elements.

As an example, the "*begin with*" semantic structure may indicate the need for a new graphic element. However, if it has already been mentioned, the user may instead want to perform an action using this element, e.g., highlighting an existing image. To address this ambiguity, the grammar uses a *context* input field to describe the scope of elements that the grammar operates on (Figure 3-1). Crosspower allows users to adjust the *context* input using lightweight interactions to achieve their desired graphic effects.

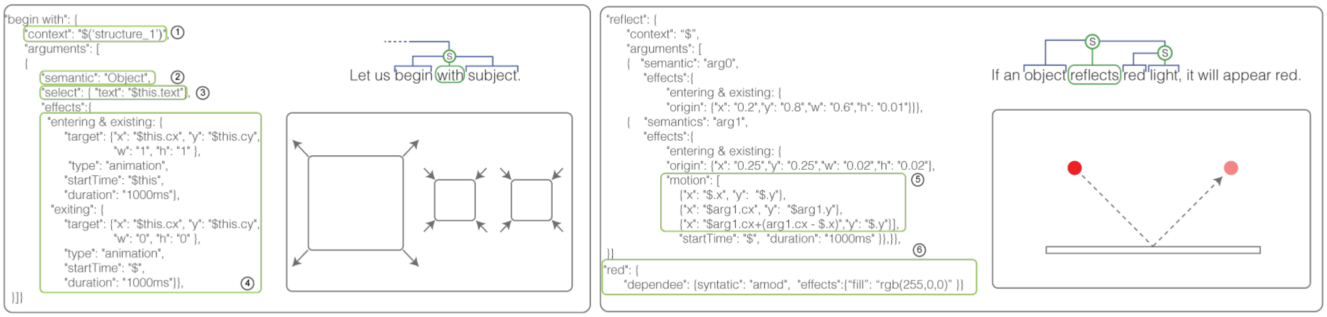
When the grammar of a semantic structure is applied (Figure 3-2), it compares the elements appearing in the arguments with those in the *context* by matching their characters and separating all elements into one of three categories (Figure 3-3):

- *Entering* elements, i.e., elements that appear in the structure but not in the *context*
- *Existing* elements, i.e., elements in the context that are referred to in the *structure*
- *Exiting* elements, i.e., elements that are in the context but not referred in the *structure*

For each category, the grammar further specifies the graphic effects based on the semantics of the structure. As an example, a "*begin with"* phrase can suggest the highlighting of the *entering* or *existing* element and optionally the blur of the *exiting* elements as well.

Specification of Graphic Effects and Behaviors

The grammar also specifies how the layouts and animations of the graphic element should represent the semantics. For example, a highlight or zoom-in effect on an image can represent the transitional action indicated by "*begin with*" (Figure 3-4). For the sentence, "*Language is the foundation of civilization*", a potential visual representation of the *foundation* relationship between *language* and *civilization* could be an image of language underneath the image of civilization to visualize that one is supporting the other.



```
"begin with": {
    "context": "$('structure_1')",
    "arguments": [
    {
        "semantic": "Object",
        "select": { "text": "$this.text"},
        "effects":{
         "entering & existing: {
             "target": {"x": "$this.cx", "y": "$this.cy",
                        "w": "1", "h": "1" },
              "type": "animation",
             "startTime": "$this",
             "duration": "1000ms"},
          "exiting": {
             "target": {"x": "$this.cx", "y": "$this.cy",
                        "w": "0", "h": "0" },
             "type": "animation",
             "startTime": "$",
             "duration": "1000ms"}},
 }]}
```

```
"reflect": {
     "context": "$",
     "arguments": [
     {   "semantic": "arg0",
         "effects":{
              "entering & existing: {
              "origin": {"x": "0.2","y": "0.8","w": "0.6","h": "0.01"}}},
     {    "semantics": "arg1",
         "effects":{
              "entering & existing: {
              "origin": {"x": "0.25","y": "0.25","w": "0.02","h": "0.02"},
              "motion": [
                  {"x": "$.x", "y":  "$.y"},
                  {"x": "$arg1.cx", "y":  "$arg1.y"},
                  {"x": "$arg1.cx+(arg1.cx - $.x)","y": "$.y"}],
                  "startTime": "$",  "duration": "1000ms" }},}},
}}
"red": {
     "dependee": {syntatic": "amod",  "effects":{"fill": "rgb(255,0,0)" }}
```

1
2
3
4
5
6
Let us begin with subject.
S
If an object reflects red light, it will appear red.
S
S

Figure 3. An example of a language-driven grammar, where (1) the context field allows the user to interactively compose structures, (2) arguments are identified based on their semantic roles, (3) selection operations based on context separate entering, existing, and exiting elements, (4) semantics are reflected using graphic effects, (5) motion paths for animations are specified, and (6) syntactic structures are specified.

The grammar supports the specification of numerical graphic attributes including position (x, y), size (width, height), motion path, and opacity, as well as attributes for animations, including the animated graphic properties, their begin and end values, as well as the start time and duration of the animation. It also allows one to reference the attributes of other graphic and language elements. This reduces the need to manually adjust graphic elements with respect to others. For example, in Figure 3-5, the motion path for "*reflect*" is described using the position attributes of the related elements, eliminating the need for the manual adjustment of the motion path when the user changes the position of the related objects.

Syntactic structures often describe other attributes of graphic elements, in addition to their layout and animation. For example, in the phrase, "*if an object reflects red light*", the word *red* modifies the color of the corresponding graphic element. Like semantic structures, the grammar also specifies corresponding graphic effects or behaviors for each constituent of a syntactic structure (Figure 3-6).

INTERACTING WITH LANGUAGE AND GRAPHICS

Crosspower provides a set of interaction techniques that allow users to quickly navigate, select, modify, and connect the linguistic and organizational structures.

Organizing, Representing, and Navigating Structures

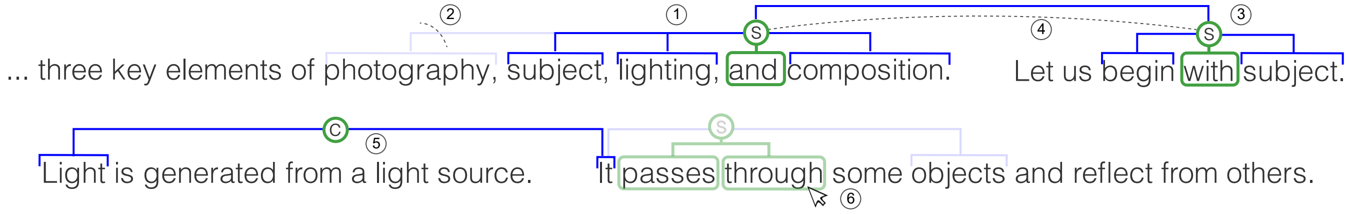
Each word in a language can be associated with many linguistic and organization structures, but not all of them indicate meaningful graphic representations. Whenever a user hovers over a word, Crosspower extracts the linguistic and organizational structures that contain the word and organizes them based on their hierarchical level, i.e., organizational structures, co-reference structures, semantic structures, syntactic structures, and then the word itself. Crosspower then suggests suitable structures to indicate the corresponding graphic structure. Crosspower prioritizes semantic structures, as they often indicate meaningful relationships that can be represented graphically (Figure 4).

The user can also navigate through the different hierarchical levels to find the one that suits their needs. Selecting the currently shown structure and will add the corresponding graphic layout or animation to the canvas.

Composing and Modifying Structures

Crosspower also allows users to compose language structures to quickly create complex graphic effects. The user can draw a connection between two structures to indicate that the elements mentioned in the previous structure will serve as the context for the grammar within the next structure. Corresponding changes in the graphic representation will then be automatically applied.

The user can also adjust the arguments used in the creation of graphic content. This can be useful if the user does not wish to visualize all the related elements in the structure or needs to fix structure extraction errors. To achieve this, they can remove existing connections or create new connections between the structure to the elements (Figure 4-2). These structural changes to the text will automatically propagate to the corresponding graphic element and vice versa.



1
2
3
4
5
6
S
S
C
S
... three key elements of photography, subject, lighting, and composition.
Let us begin with subject.
Light is generated from a light source.
It passes through some objects and reflect from others.

Figure 4. An example of a (1) syntactic conjunction structure, where (2) the mis-extraction of the syntactic structure can be fixed by removing the unwanted element. (3) A semantic structure with its semantic arguments. (4) The use of a previous structure as context by connecting the two structures. (5) A coreference structure. (6) Crosspower suggesting a suitable structure as the user hovers over the words. The icon on top of each structure indicates the type of structure, with S indicating both syntactic and semantic structures and C indicating coreference structures.

In cases where an NLP toolkit fails to recognize coreference structures that users wish to leverage, Crosspower allows users to connect linguistic elements to create new coreference structures to add graphic effects to existing elements instead of creating new ones.

Visualizing and Editing Graphic Structures

Once the user confirms that they want to visualize certain structures, a default graphic representation is added to the canvas. For entering elements, Crosspower utilizes the Google Image Search to query images with corresponding text as the query and displays the first returned image on the canvas. In some cases, the user may want to select a different image and Crosspower allows the user to browse all the returned search results in-situ, without context switching. The user can also change the search query of each graphic element to query new sets of images.

If a user wishes to adjust their search by constraining queries with the same new keywords (if they are related to the same domain or concept), Crosspower enables the user to propagate the addition or removal of keywords to all the structured elements to consistently apply the adjustment.

Flexible Graphic Representations

The user may wish to use various graphic elements, such as image, shapes, or text to represent underlying concepts. Crosspower allows the users to flexibly combine and replace graphic representations to match their own design aesthetics. The user can select the graphic structure and then toggle amongst the different representations to switch the representation or select and combine multiple representations (Figure 5). This allows users to quickly experiment with different visual effects using different graphic representations. In some cases, there can be multiple graphic effects associated with one linguistic structure. Crosspower displays all other graphic effects to the right of the interface so that the user can browse and select the one that suits their needs.

Text Selection and Lists as Structures

Users often visualize text in a script directly on the canvas to highlight important messages, communicate inherent textual information, or for labelling purposes. To support such needs, Crosspower enables users to select text and transform it into self-defined linguistic structures. The use of these structures automatically creates text elements on the canvas. With a time-aligned script, Crosspower can automatically create a 'text revealing' animation where each word will appear the moment it is narrated in a video. Similarly, Crosspower supports users in directly converting text lists in a script to graphic lists, where each list item appears based on the timing in a narration (Figure 6-4).

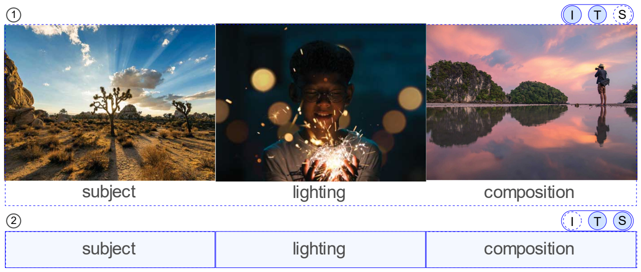
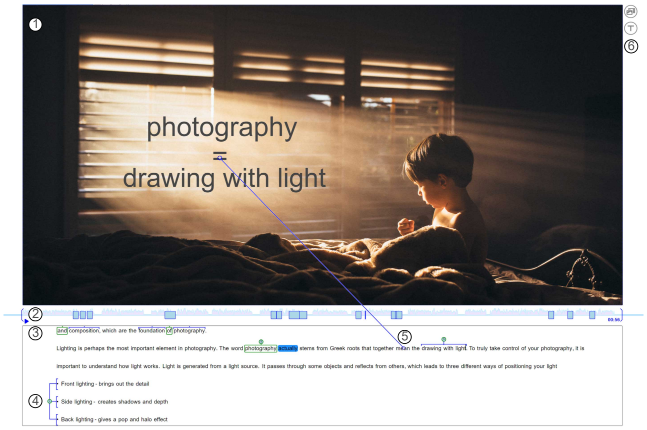


Figure 5. Flexible composition of graphical representation. With the same language-grammar, the user can combine images, text, and shapes, to create different graphic effects.



and composition, which are the foundation of photography.

Lighting is perhaps the most important element in photography. The word photography actually stems from Greek roots that together mean the drawing with light. To truly take control of your photography, it is important to understand how light works. Light is generated from a light source. It passes through some objects and reflects from others, which leads to three different ways of positioning your light

- Front lighting - brings out the detail
- Side lighting - creates shadows and depth
- Back lighting - gives a pop and halo effect

Figure 6. The Crosspower interface: with 1) the main canvas, 2) a timeline with a set of added animation, 3) script section, 4) a list structure being used to create a graphic list, 5) connecting a graphic element to a linguistic element, and 6) basic editing tools.

Bi-directional Mapping

In cases where none of the provided graphic styles suits a user's needs or the user wishes to begin with their own creations, they can manually create a desired style using the basic editing operations provided with Crosspower, such as adding new images and text to the canvas, or configuring their size, position, or animations.

Once a graphic effect is created, the user can connect the graphic object with its corresponding language element in the script (Figure 6-5). This allows the user to align the timing of the animation to the narration, but also allows Crosspower to extract spatial layouts and animation properties to form a new graphic representation for the underlying linguistic structure.

CROSSPOWER WORKFLOW

To demonstrate the utility of Crosspower, we will walk through an example workflow by following Hayley, a professional photographer and YouTuber, who regularly creates and posts videos. Today she is starting to work on a video that introduces basic concepts in photography.

As always, Hayley first works on her video script to determine the content she will cover in her video. Once her script is done, she then records a voice-over of the script and begins to create graphic content with Crosspower.

"I will talk about three key elements of photography, subjects, lighting, and composition." For this opening sentence, she would like to have an overview animation that shows a representative image for each element, one by one. She can directly create corresponding graphic elements and their effects by leveraging the conjunction structure in the

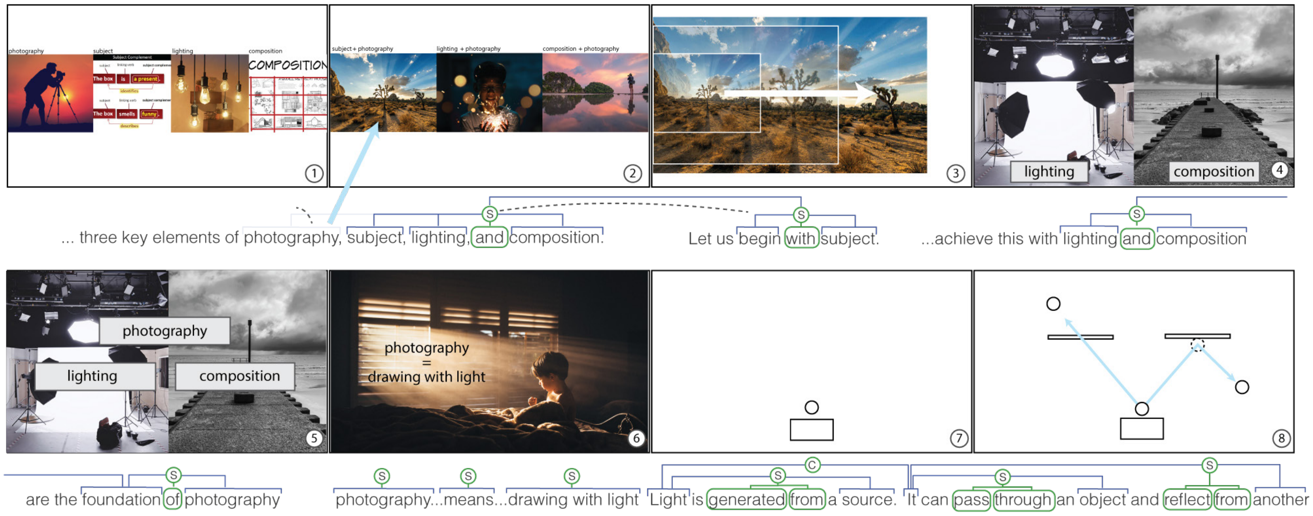

photography
subject
Subject Complement
subject
linking verb
subject complement
The box
is
a present
identifies
The box
smells
funny
describes
lighting
composition
COMPOSITION
1
subject + photography
lighting + photography
composition + photography
2
3
lighting
composition
4
S
... three key elements of photography, subject, lighting, and composition.
Let us begin with subject.
...achieve this with lighting and composition
photography
lighting
composition
5
photography
=
drawing with light
6
7
8
C
are the foundation of photography
photography...means...drawing with light
Light is generated from a source.
It can pass through an object and reflect from another

Figure 7. Example Crosspower Workflow where 1) a conjunction structure is added, 2) the structure is modified, 3) a connection is made to the "*begin with*" structure, 4) shape and text are used to represent the structure, 5) a "*foundation*" structure is created, 6) text selections are used as structures, 7) a "*generated from*" effect is created, and 8) an animation is created using the "*pass through*", "*reflect from*" structures, and the coreference structure between "*light*" and "*it*".

text. However, the underlying dependency parser makes an error and extracts "*photography*" as one of the elements in the structure (Figure 7-1). She simply crosses the word "*photography*" out in the structure and the graphic elements are automatically adjusted. Crosspower uses the text of the elements as search queries to automatically find images online. However, the returned images are not ideal and she realizes she needs to constrain the queries, so she drags and drops the word "*photography*" from the script to the canvas to use it as an additional search keyword (Figure 7-2).

"*Let's begin with subject*". Here Hayley wants an expansion animation of the subject image. She can directly select the animation indicated by "*begin with*" using Crosspower. The system then creates a grow animation for a new image element. This is because the underlying NLP toolkit fails to recognize the coreference of "*subject*". She easily fixes this error by indicating that the previous conjunction structure should be used as the context for "*begin with*". This not only allows for the creation of the "*subject*" grow effect, but also the shrink effect of the other element (Figure 7-3).

"*You can achieve this with lighting and composition, which are the foundation of photography.*" Here, she can create an effect where lighting and composition are represented in rectangular textboxes underneath the *photography* textbox, representing the concept of foundation. A coreference structure will also need to be created between "*lighting and composition*" and "*foundation*" (Figure 7-5).

"*The word photography actually stems from Greek roots that mean drawing with light*". Here, she would like to create a text effect with "*photography*" and "*drawing with light*", where the words reveal themselves one by one. She then selects the text and creates the corresponding graphic text elements. She can also manually creates a "=" text element on the canvas and maps it to "*mean*" in the script to leverage its temporal information (Figure 7-6).

"*Light is generated from a light source. It passes through some objects and reflects from others.*" Here, a sequence of animations indicated by "*generate*", "*pass*", and "*reflect*" can be chained together thanks to their coreference structures and the pre-defined animations associated to the semantic structures. She can also further customize the motion paths associated with these animations (Figure 7-7).

The above example demonstrates how a user can directly select, define, navigate, modify, and combine various linguistic structures to quickly create corresponding graphic effects. This supports a user in expressing high-level design goals rather than performing tedious low-level operations.

EXPERT EVALUATION

To validate that extending direct manipulation to structures in language and leveraging the correspondences between graphics and linguistics can enable the flexible and direct creation of graphic content, and to gain feedback about the usefulness and effectiveness of the techniques introduced in Crosspower, an expert evaluation study was conducted.

Participants

Six professional video, animation, and presentation creators were recruited online to evaluate Crosspower in a remote-participation study (2 female, aged 28 – 42 years). All participants have at least 7 years' experience creating videos, animation, or presentations. Participants were requested to provide their professional evaluation on whether and how Crosspower will be useful for their content creation process. Participants received $60 (USD) for the approximately 90-minute session.

Apparatus

To facilitate remote participation, Crosspower was run on a Windows PC which participants were able to directly interact with through TeamViewer. Video conferencing was used to and communicate with participants.

Procedure

Each expert review session included the following phases:

Introduction and Training (25 minutes)

The experiment first introduced the underlying concepts of Crosspower. Then, the experimenter performed the interaction techniques and described them verbally. Participants then were then asked to perform the interaction techniques and seek help when necessary.

Creation Exercise (20 minutes)

The participants were then asked to create and iterate on the graphic content for a 205-word script provided by the experimenter, which lent itself to many of the implemented interaction techniques. This task was designed to ensure that participants got enough practice using Crosspower and for the research team to observe their learning process.

Freeform Exploration (20 minutes)

Participants were then asked to create the graphic content for a segment of a video or presentation script (200-250 words) that they had previously used, which the experimenter requested them to bring to the study.

Questionnaire and Exit Interview (20 minutes)

Participants then completed a questionnaire about Crosspower, probing the usefulness and usability of the interaction techniques using a 7-point Likert scale (1 – Strongly Disagree, 7 - Strongly Agree). The experimenter then conducted a semi-structured interview to further collect feedback about the utility of the interaction techniques and the workflow when using Crosspower.

Results

We report on the results of the expert evaluation pertaining to the utility of language structures, the new workflows enabled by Crosspower, suitable content domains.

Utility of Language Structures

Participants were asked to rate the usefulness of each of the techniques (Figure 8). The results indicated that the various techniques to interact with the language structures were useful and desirable. All participants responded positively that the use of language structures allowed them to quickly (4/6 strongly agree, 2/6 agree) and flexibly (4/6 strongly agree, 2/6 agree) create graphic content.

Participants also responded favorably to the composition of language structures, as they removed the “*painful work*” (P4) and allowed them to “*quickly build something pretty complex*” (P5). Being able to modify structures, as well as select and combine different representations, was also preferred by participants as it enabled “*control over the provided templates”* (P2).

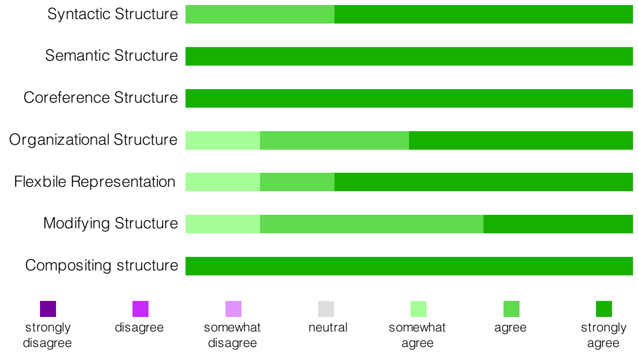


Figure 8. Likert-scale responses to “This technique was useful in my animation creation tasks…”

Language Structures vs. Existing Practices

Participants found that “*the use of language structures fits my prior workflows of creating graphic content (i.e., video, animation, or presentation)*” (6 - strongly agree). “*The content and message are most important*” (P1), and it is important to “*make sure graphics match the content*” (P3), as “*you want to use the graphics to help the audience to understand the content not to confuse them*” (P3).

Leveraging language structures also enabled new workflows, where participants were able to focus on exploring what content they wanted to represent graphicly rather than on how to create the graphics, e.g., “*I felt like I was mostly focusing on the script and less on how to make the effects, but still I was able to create good effects at the end* ” (P5) and “*you can quickly throw together a decent deck of slides with this*” (P1).

Participants acknowledged the strength of encapsulating several animations into a language structure, i.e., “*the animations provided in most existing tools are very basic …. you need to know the big transition you want, and then figure out how to achieve that with the basic animations … and this is not easy, especially when I first started [in this domain]*” (P1). With Crosspower, they could directly see the potential animations that “*represent the messages*” (P5). Moreover, participants perceived the language structures as a suitable way to organize or index the templates “*when I search templates, I need to use exact names of the effects to get good results, but sometimes I don’t know what effects are good, it would be great if I can search with the content itself and see what’s out there* ” (P2).

When asked to compare language-driven templates to other templates they have used, participants appreciated the ability to modify the underlying language structure to adjust the graphic templates, as the “*adaptable templates”* allowed them to “*easily turn the templates into what I [they] want”* (P4). In comparison, they often need to “*do a lot of tweaks for the templates I [they] got online”* (P4).

Suitable Domains

Participants suggested several types of graphic content that could be easily created with Crosspower, including technical presentations and informational videos (e.g. explainer videos, video essays, or infographic videos), which often utilize animation and graphics to facilitate content comprehension and can benefit from the correspondence between linguistics and graphics.

Participants also commented that content that is either too formal or informal may not be well suited for Crosspower. Formal content such as motion graphics often requires precise specification and is “*more to dazzle the audience*”

(P5) rather than communicating meaningful information. On the other hand, creative and artistic expressions, such as inspirational talks or poetry, often consist of abstract, ambiguous, or emotional words and phrases that may not have direct linguistic-graphic correspondences and be better accompanied by specific and well-chosen images (P1).

Limitation and Discussion

The results from the expert evaluation show that leveraging the linguistic-graphic mappings could reduce the manual effort encountered when creating graphic content, but also suggest limitations and opportunities for improvement.

Erroneous Natural Language Processing

Crosspower builds upon linguistic structures provided by NLP toolkits, which contain errors occasionally, including failing to extract semantic and coreference structures as well as erroneous syntactic and semantic parsing. Crosspower does not support the correction of parsing errors or the specification of missing semantic structures. All participants felt confused when encountering such errors and had to resort to manual creation. While we expect the mitigation of such problems as NLP techniques become increasingly powerful, an alternative might be to enable users to specify desired changes which can be propagated to underlying NLP modules for error correction.

Complex Linguistic Structures

Similar confusion was also found when participants were shown complex semantic structures that contain multiple hierarchical levels or many semantic arguments, when they were only interested in parts of structures. Participants commented that they were *overwhelmed by the complexity* (P1, 3), and had to spend time understanding the structures and deciding how to utilize them. This is perhaps because the linguistic structures provided by NLP toolkits do not directly match users' expectation. This can be addressed by progressively disclosing the semantic structures and arguments based on context and in a representation that matches user's mental models.

While the expert evaluation demonstrates considerable promise of language-oriented authoring, our user evaluation is preliminary and can benefit from in-depth investigation into how the concept can be applied to graphic content of different types with users of different levels of expertise.

FUTURE WORK

A Universal Linguistic-Graphic Dictionary

An immediate next step is to collect a large amount of language-driven graphic effects. A repository of graphic effects will increase the expressive power of Crosspower, but also allow for the further exploration of how to suggest the most suitable graphic effect for a language structure that fits into a holistic visual style. This collected repository will also contribute to efforts to construct a complete a linguistic-graphic dictionary. Prior efforts such as ImageNet [13] focused on the construction between *nouns* to images of real-world objects, whereas Crosspower has focused on the dynamic graphic actions mostly indicated by *verbs*. Today, an increasing number of graphic-rich videos such as explainer and infographic videos are published online, which use graphics, animations, and narration to explain concepts with a compelling storytelling experience. These videos contain rich linguistic-graphic mappings that we will collect and share with the community.

Interactive Scripting Graphic Content

Crosspower currently does not support the interactive experimentation between linguistic and graphic structures. While a user is free to edit the natural language input and interact with the linguistic structures, Crosspower may not be able to provide the desired graphic effects due to the limited pre-defined linguistic-graphic mappings. The rich linguistic-graphic dictionary we set to establish will allow us to explore the interactive creation and modification of both linguistic and graphic content. We envision this will enable new ways of creating graphic content as the users dynamically experiment both the linguistic and graphic expression. It will also be interesting to explore the combination of explicit scripts or markup languages together with natural language input to enable for the quick and flexile composition of graphic content.

From Written to Spoken Language

While Crosspower has focused on leveraging structures in written language, the use of linguistic structures can also be directly applied to spoken language. This can be useful for generating visual aids during conversations in augmented reality or other forms of shared displays. Besides rich linguistic structures, spoken language uses acoustic signals such as pitch, tone, and stress to convey meaning and sentiment, which could be useful to infer graphic styles.

Challenges with Creative and Artistic Expression

While linguistic-graphic mappings allow content creators to articulate high-level design goals, they may not be abstracted enough for creative and artistic language expressions, which often consist of words and phrases that are abstract, ambiguous, or emotional. Such high-level semantics often do not have clear and direct graphic correspondences and require creative composition of graphic effects. We seek to identify, distill, and leverage such higher-level design knowledge and creativity to facilitate the design of compelling graphic content.

CONCLUSION

We present a systematic exploration of language-oriented authoring which bridges linguistics with graphics through the identification, graphic specification, and interaction of various structures in written language. The research prototype, Crosspower, enables users to directly navigate, select, modify, and compose linguistic structures to indicate high-level design goals rather than forcing users to perform tedious low-level editing operations. Demonstrated through expert evaluation, Crosspower enabled content creators to create and customize graphic content directly and flexibly.

REFERENCES

[1] Maneesh Agrawala, Wilmot Li, and Floraine Berthouzoz. 2011. Design principles for visual communication. Commun. ACM 54, 4 (April 2011), 60-69. DOI: https://doi.org/10.1145/1924421.1924439

[2] Olga Babko-Malaya. (2005). Propbank annotation guidelines. URL: http://verbs. colorado. edu.

[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pp. 86-90. Association for Computational Linguistics, 1998.

[4] Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI ’00). Association for Computing Machinery, New York, NY, USA, 446–453.

DOI: https://doi.org/10.1145/332040.332473

[5] Michel Beaudouin-Lafon and Wendy E. Mackay. 2000. Reification, polymorphism and reuse: three principles for designing visual interfaces. In Proceedings of the working conference on Advanced visual interfaces (AVI ’00). Association for Computing Machinery, New York, NY, USA, 102–109. DOI: https://doi.org/10.1145/345513.345267

[6] Benjamin B. Bederson, James D. Hollan, Allison Druin, Jason Stewart, David Rogers, and David Proft. "Local tools: An alternative to tool palettes." In Proceedings of the 9th annual ACM symposium on User interface software and technology, pp. 169-170. 1996.

[7] Richard A. Bolt. 1980. “Put-that-there”: Voice and gesture at the graphics interface. In Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH '80). ACM, New York, NY, USA, 262-270. DOI: http://dx.doi.org/10.1145/800250.807503

[8] Angel Chang, Manolis Savva, and Christopher D. Manning. "Learning spatial knowledge for text to 3D scene generation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[9] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: content creation with semantic attributes. In Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST ’13). Association for Computing Machinery, New York, NY, USA, 193–202. DOI: https://doi.org/10.1145/2501988.2502008

[10] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. QuickSet: multimodal interaction for distributed applications. In Proceedings of the fifth ACM international conference on Multimedia (MULTIMEDIA '97). ACM, New York, NY, USA, 31-40. DOI: http://dx.doi.org/10.1145/266180.266328

[11] Bob Coyne and Richard Sproat. 2001. WordsEye: an automatic text-to-scene conversion system. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques (SIGGRAPH '01). ACM, New York, NY, USA, 487-496. DOI: https://doi.org/10.1145/383259.383316

[12] Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang. "Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements." IEEE transactions on visualization and computer graphics 26, no. 1 (2019): 906-916.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.

[14] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’08). Association for Computing Machinery, New York, NY, USA, 237–246. DOI: https://doi.org/10.1145/1357054.1357096

[15] Marianela Ciolfi Felice, Nolwenn Maudet, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2016. Beyond Snapping: Persistent, Tweakable Alignment and Distribution with StickyLines. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST ’16). Association for Computing Machinery, New York, NY, USA, 133–144. DOI: https://doi.org/10.1145/2984511.2984577

[16] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based editing of talking-head video. ACM Trans. Graph. 38, 4, Article 68 (July 2019), 14 pages. DOI: https://doi.org/10.1145/3306346.3323028

[17] Bill Gates. 1996. Content is king. Retrieved October, 29, p.2017.

[18] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST ’15). Association for Computing Machinery,

New York, NY, USA, 489–500. DOI: https://doi.org/10.1145/2807442.2807478

[19] Google NLP. 2020. https://cloud.google.com/natural-language/

[20] David G. Hays "Dependency theory: A formalism and some observations." Language 40, no. 4 (1964): 511-525.

[21] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7986–7994.

[22] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J. Mysore. 2019. B-Script: Transcript-based B-roll Video Editing with Recommendations. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, New York, NY, USA, Paper 81, 11 pages. DOI: https://doi.org/10.1145/3290605.3300311

[23] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03), 17-24. http://dx.doi.org/10.1145/642611.642616.

[24] Takeo Igarashi and John F. Hughes. 2001. Voice as sound: using non-verbal voice input for interactive control. In Proceedings of the 14th annual ACM symposium on User interface software and technology (UIST '01). Association for Computing Machinery, New York, NY, USA, 155–156. DOI: https://doi.org/10.1145/502348.502372

[25] Dhiraj Joshi, James Z Wang, and Jia Li. 2004. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In Proceedings of the 6thACM SIGMM international workshop on Multimedia information retrieval. ACM, 119–126.

[26] Murat Kalender, M. Tolga Eren, Zonghuan Wu, Ozgun Cirakman, Sezer Kutluk, Gunay Gultekin, and Emin Erkan Korkmaz. 2018. Videolization: knowledge graph based automated video generation from web content. Multimedia Tools and Applications 77, 1 (01 Jan 2018), 567–595. DOI: http://dx.doi.org/10.1007/s11042-016-4275-4

[27] Chen, Kevin, Christopher B. Choy, Manolis Savva, Angel X. Chang, Thomas Funkhouser, and Silvio Savarese. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In Asian Conference on Computer Vision, pp. 100-116. Springer, Cham, 2018.

[28] Daniel Khashabi, Mark Sammons, Ben Zhou, Tom Redman, Christos Christodoulopoulos, Vivek Srikumar, Nick Rizzolo et al. "Cogcompnlp: Your swiss army knife for nlp." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[29] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015.Ranking and retrieval of image sequences from multiple paragraph queries. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.1993–2001.

[30] Yea-Seul Kim, Mira Dontcheva, Eytan Adar, and Jessica Hullman. 2019. Vocal Shortcuts for Creative Experts. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 332, 1–14. DOI: https://doi.org/10.1145/3290605.3300562

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123, no. 1 (2017): 32-73.

[32] Gierad P. Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: a multimodal interface for image editing. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). Association for Computing Machinery, New York, NY, USA, 2185–2194. DOI: https://doi.org/10.1145/2470654.2481301

[33] Mackenzie Leake, Hijung Valentina Shin, Joy Kim and Maneesh Agrawala. 2020. Generating Audio-Visual Slideshows from Text Articles Using Word Concreteness. ACM Human Factors in Computing Systems (CHI), Apr 2020. To Appear.

[34] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP natural language processing toolkit." In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60. 2014.

[35] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. 2017. Attentive semantic video generation using captions. In Proceedings of the IEEE International Conference on Computer Vision.1426–1434.

[36] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. "The NomBank project: An interim report." In Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004, pp. 24-31. 2004.

[37] George A. Miller. "WordNet: a lexical database for English." Communications of the ACM 38, no. 11 (1995): 39-41.

[38] Sharon Oviatt. 1999. Ten myths of multimodal interaction. Commun. ACM 42, 11 (November 1999), 74–81. DOI: https://doi.org/10.1145/319382.319398

[39] Ken Perlin. "Future Reality: How emerging technologies will change language itself." IEEE computer graphics and applications 36, no. 3 (2016): 84-89.

[40] Martha Palmer, Daniel Gildea, and Paul Kingsbury. "The proposition bank: An annotated corpus of semantic roles." Computational linguistics 31, no. 1 (2005): 71-106.

[41] Steve Rubin, Floraine Berthouzoz, Gautham Mysore, Wilmot Li, and Maneesh Agrawala. 2012. UnderScore: Musical Underlays for Audio Stories. In Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12). ACM, New York, NY, USA, 359–366. DOI: http://dx.doi.org/10.1145/2380116.2380163

[42] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based tools for editing audio stories. In Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST '13). ACM, New York, NY, USA, 113-122. DOI: https://doi.org/10.1145/2501988.2501993

[43] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16). ACM, New York, NY, USA, 365-377. DOI: https://doi.org/10.1145/2984511.2984588

[44] Ben Shneiderman. (1982). The future of interactive systems and the emergence of' direct manipulation. Behavior and Information Technology, 1, 237-256.

[45] Hijung Valentina Shin, Wilmot Li, and Frédo Durand. 2016. Dynamic Authoring of Audio with Linked Scripts. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16). ACM, New York, NY, USA, 509–516. DOI: http://dx.doi.org/10.1145/2984511.2984561

[46] Vivek Srikumar and Dan Roth. "Modeling semantic relations expressed by prepositions." Transactions of the Association for Computational Linguistics 1 (2013): 231-242.

[47] Arjun Srinivasan and John Stasko. "Orko: Facilitating multimodal interaction for visual exploration and analysis of networks." IEEE transactions on visualization and computer graphics 24, no. 1 (2017): 511-521.

[48] Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphic Communication System*. Ph.D Dissertation. MIT, Cambridge, MA.

[49] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. QuickCut: An Interactive Toolfor Editing Narrated Video. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16). ACM, New York, NY, USA,497–507. DOI: http://dx.doi.org/10.1145/2984511.2984569

[50] Barbara Tversky, Julie Bauer Morrison, and Mireille Bétrancourt (2002). Animation: can it facilitate?. International journal of human-computer studies, 57(4), 247-262.

[51] Edward R. Tufte. 1997. Visual Explanations: Images and Quantities, Evidence and Narrative. Graphics Press, Cheshire, CT, USA.

[52] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. 2018. Deep convolutional priors for indoor scene synthesis. ACM Trans. Graph. 37, 4, Article 70 (July 2018), 14 pages. DOI: https://doi.org/10.1145/3197517.3201362

[53] Terry Winograd. 1972. Understanding Natural Language. Academic Press, Inc., Orlando, FL, USA.

[54] Haijun Xia, Bruno Araujo, and Daniel Wigdor. 2017. Collection Objects: Enabling Fluid Formation and Manipulation of Aggregate Selections. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). Association for Computing Machinery, New York, NY, USA, 5592–5604. DOI: https://doi.org/10.1145/3025453.3025554

[55] Haijun Xia, Bruno Araujo, Tovi Grossman, and Daniel Wigdor. 2016. Object-Oriented Drawing. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 4610–4621. DOI: https://doi.org/10.1145/2858036.2858075

[56] Haijun Xia, Jennifer Jacobs, Maneesh Agrawala. Crosscast: Adding Visuals to Audio Travel Podcasts. In Proceedings of the 33rd annual ACM symposium on User interface software and technology (UIST '20). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3379337.3415882

[57] Haijun Xia, Nathalie Henry Riche, Fanny Chevalier, Bruno De Araujo, and Daniel Wigdor. 2018. DataInk: Direct and Creative Data-Oriented Drawing. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 223, 1–13. DOI: https://doi.org/10.1145/3173574.3173797