

What's Next?: Emergent Storytelling from Video Collections

Edward Yu-Te Shen
MIT Media Laboratory
edward@media.mit.edu

Henry Lieberman
MIT Media Laboratory
lieber@media.mit.edu

Glorianna Davenport
MIT Media Laboratory
gid@media.mit.edu



Figure 1: Three steps in the edit-by-recommendation functionality

ABSTRACT

In the world of visual storytelling, narrative development relies on a particular temporal ordering of shots and sequences and scenes. Rarely is this ordering cast in stone. Rather, the particular ordering of a story reflects a myriad of interdependent decisions about the interplay of structure, narrative arc and character development. For storytellers, particularly those developing their narratives from large documentary archives, it would be helpful to have a visualization system partnered with them to present suggestions for the most compelling story path.

We present *Storied Navigation*, a video editing system that helps authors compose a sequence of scenes that tell a story, by selecting from a corpus of annotated clips. The clips are annotated in unrestricted natural language. Authors can also type a story in unrestricted English, and the system finds possibilities for clips that best match high-level elements of the story. Beyond simple keyword matching, these elements can include the characters, emotions, themes, and story structure. Authors can also interactively replace existing scenes or predict the next scene to continue a story, based on these characteristics. *Storied Navigation* gives the author the feel of *brainstorming about the story* rather than simply *editing the media*.

ACM Classification Keywords

I.3.6 Methodology and Techniques — Interaction Techniques; H.5.2 Information Interfaces and Presentation — Natural Language, Theory and Methods

Author Keywords

Interactive storytelling, emergent storytelling, storied navigation

INTRODUCTION

In the world of visual storytelling, narrative development

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2009, April 4–9, 2009, Boston, MA, USA.

Copyright 2009 ACM 978-1-60558-246-7/09/04...\$5.00.

relies on a particular temporal ordering of shots and sequences and scenes. This ordering reflects a myriad of decisions about the interplay of structure, narrative arc and character development, particularly for documentary storytelling, i.e. producing documentary films or home videos. These decisions are interdependent and often difficult to make at the time individual scenes are generated, and the storyteller is often faced with the problem of how to choose amongst a large set of alternatives for presenting the story. That is, even if an author initially thinks her story has been fully specified, it is common for her to change her mind while actually watching the material or transition between scenes in the context of the whole story. It would be helpful, thus, to have a visualization system partnered with her to present the most compelling story path, or to suggest new possibilities that she had not thought of. In this case, a coherent and compelling story can emerge when a story author considers selecting scenes from a large *corpus* of available material, which we call *emergent storytelling*.

This paper presents *Storied Navigation*, a novel video editing system that engages the storyteller in an interactive process in which she retrieves prerecorded videos by typing free-text stories, and composes the temporal ordering of shots, sequences, and scenes. It allows users to construct a narrative in order to substantiate a specified story goal, or simply to explore the materials with no preconceived notions about what story might be told, and let the exploration itself suggest a possible story. The goal of the system is to emulate ways of thinking about the material that mirror human storytelling strategies. For instance, one rendition might return a set of sequences that show a temporal progression that matches a chronological progression of events, while another rendition might emphasize the typical behavior of a protagonist.

The human capacity for storytelling develops in parallel with human capacity for language. While other video-retrieval and recommendation systems work simply by matching keywords, we believe the system must be able to understand and associate the basic meaning of a scene with the meaning found in other parts of the story. We use state-of-the-art natural language processing to achieve at least a partial understanding of human descriptions of scenes and story. Further the system must be able to relate

what happens in a scene to some basic concepts about story structure. The theories behind the functionalities of this system are derived from an analysis of documentary filmmaking, and are implemented by applying commonsense reasoning technology from AI, because of its unique capability for emotion sensing and analogy making.

The system works in two phases. First, an *annotation* phase describes individual scenes, either in unrestricted natural language, or by a form-filling interface where the user can explicitly describe certain attributes, such as characters, emotions or themes. Annotation of scenes could also be obtained from transcription of dialog, automatic video/audio analysis, or other sources. Second, a *composition* phase composes a temporal sequence out of a corpus of individual scenes. The system provides several recommendation functionalities in this phase, including *Edit-by-Typing*, where the author can type a story and the system recommends clips appropriate for each element of the story; and *Edit-by-Recommendation*, where she asks “What’s Next?” and the system returns a set of candidates for the next clip, by virtue of continuing the story structure. Given a sequence, the author can also ask for “Similar Alternatives”, to retrieve sequences that express the same storytelling intent. The retrieval and recommendation facilities are integrated in with a timeline-based video editor that allows immediate preview of selections at any time. This shows how the facilities can be integrated with commercial editors such as Final Cut or Adobe Premiere.

A fundamental problem that Storied Navigation tries to address is the mismatch in intent between the kinds of descriptions that typically appear in scene annotation, and those appear in story descriptions while the storyteller is developing the narrative. Scene annotations typically tend to be very concrete, describing the individual characters and

actions, the “who-what-where-when” of the scene. Story descriptions, on the other hand, tend to discuss things like plot elements, emotions, roles and intentions of characters, reasons for doing things, etc. Our Commonsense knowledge base and reasoning technology is intended to bridge the gap between a story description, such as, “A mentoring relationship between a teacher and a student”, and a scene description, such as “Jacques helps Gustave understand why he must cooperate with Duffin.”

We also present our user studies and discussion. The first study investigates whether the system is useful to authors in creating interesting and cogent story threads. The second tests whether the system is helpful in familiarizing users with the material. This is an important criterion as the annotator and storyteller are not the same person, or significant time has elapsed between material collection and usage. Finally, two properties of experience brought by the system are discussed, including “the accompanied editor” and “the democratic storytelling.”

THE “WAYS TO THINK”

This section presents how we derived the set of “ways to think” that human storytellers exploit in narrative development by analyzing a film. The term “ways to think” is used by Minsky [18] to denote high-level problem solving strategies that connect a person’s goals with the methods that can achieve them. Then, we conclude a corresponding set of story features that we think today’s computers may be able to process.

Film structure analysis has been a practice that human story consultants do to help filmmakers make better stories [21], and we believe it can “help” computers to “make better

Table 1: Transition Analysis in the Example Film






Name	Thumbnail	Sequence Description	The Reason Why the Sequence is “Here”
Lab scene 2		Herr demonstrating the technology on his own legs and answering a few questions.	It shares a similar image (somebody wearing artificial limbs walking on the stairs), and similar context (lab, prosthesis research) with the previous clip. It also shows that Herr himself is an amputee, and more importantly, because of some accident, which has to be introduced here for the rest of the material to follow
Ambition		Herr talking about his plan for commercialization and the future goal while answering a visitor’s questions.	Again, it shows similar context with the previous clip. It also shows the <i>ambition</i> and <i>goal</i> that Herr has, which will lead to and resonate with the <i>reason</i> why he is doing this in the forthcoming clips.
Courage in career decisions 1		Herr’s student Sam relates how Herr convinced him to change his field	It shares and strengthens the idea that Herr wants to make the prosthesis for other people’s good. That Herr encourages other people to change directions because of “what’s right” also starts to resonate with the rest of the film from here.
Courage in career decisions 2		Herr and another student conversing in the car about the student’s career	It is analogous to the previous scene in that both scenes give a story about Herr advising a student in a career decision making situation. As a real-life event, it also makes the previous interview livelier in front of the audience.
Surprising Fact		Herr running around a lake on a pair of artificial legs.	While the rhythm of the story gradually gets calm and slow, it surprises the audience with a stunning shot. It also prepares the appearance of the next coming clip by setting the context and lifting the emotion of the story

Table 2: “Ways to Think” for placing a sequence in a story

- a. This video sequence reveals some information that has to be delivered to the audience at this exact point,
- b. This video sequence exhibits an opening, ending, or other kinds of function to the whole structure,
- c. This video sequence establishes a new context for a new story part to follow,
- d. This video sequence strengthens, extends the story, or elevates the current story to a different level,
- e. This video sequence shares analogous or related concepts (relationships, attitude, interaction, etc) with the previous sequence,
- f. This video sequence is continued from the previous sequence in terms of time, location, character, event

stories” too. The example film we choose is a 15-minute documentary film, entitled "Life. Research.", directed by the first author. It was screened at Plymouth Independent Film Festival and awarded first place at the MIT Media Spectacle event, both in 2007. It was a story cut from a collection of 12-hours of raw footage, about Hugh Herr, both a double amputee and a prestigious professor working on artificial limbs. Step by step, the story reveals the difficulties that Herr faced, his interaction with other people, and finally his personality and philosophy for life.

To investigate why each clip is placed at its current position and how the transition between two consecutive clips makes sense to the audience, Table 1 lists the inter-relationships within part of the film. Each row represents, from left to right, a sequence’s name, thumbnail, brief description, and the reason for its chronological position in the film. If we generalize the justifications for each choice of arrangement, a set of criteria for choosing a sequence at a particular point, or the “ways to think” [18] in storytelling, can be derived, as listed in Table 2. From these, we derived a set of “story features” for a video sequence that may help a system recommend adjacent sequences, including: 1) characters, 2) interactions between its characters, 3) characters’ emotions, 4) time or date, 5) location, 6) theme (e.g. “Victoria’s learning problem”), and 7) the story structural role (e.g. “beginning”, “ending”, etc) this video sequence may play in the story of which it is a part.

As these story features are concrete enough that they can be reasonably understood in many cases by our natural language processing, we can feasibly support this kind of "storied thinking" in composing stories from clips.

COMPOSING STORIES IN STORIED NAVIGATION

We now describe the process of constructing a story from a set of clips, assuming each clip has been annotated with a few sentences describing who is in the scene and what takes place in it. Later, we return to the description of the interface for annotating clips. To start a story, the user can either drag a video clip from the text area above the timeline, or “start with something arbitrary”. Storied Navigation supports “storytelling by typing” -- the user can input a piece of story text and the system will try to compose a string of videos that it finds most semantically similar to the story text.

Figure 2 shows the storytelling interface, with the parsed result of the story description, “Gustave is arguing with the

teacher because he wants to make his own dance, not just being a ‘puppet’.” The system responds with two video clips, which make the most suitable video stream in response to the story description. If the user wants more, he/she can click on the video clips and choose “show more search results” from the menu. The video player automatically plays the returned video clips from the first one, and the popup yellow box shows the annotated story description for the video clip that the system is playing back or the mouse cursor points to. In both the input text area and the popup boxes, the matched characters, actions, and emotional terms will be highlighted in boldface, so the users can tell how the clips are selected. In this example, “gustave” is matched with “Gustave”, “arguing” is matched with “upset”, and “teacher” is matched with “Jacques”, which are all correct matches.

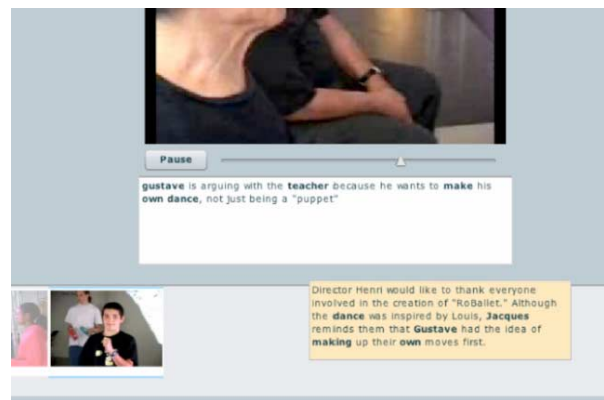


Figure 2. Composition interface

If a user is not familiar with the video corpus, he/she can input an unspecific sentence like, “someone is happy”, or “people are discussing something”, and the system will try to find the best result it can. Thus, a user can gradually “enter the story world”, familiarizing him or herself with the corpus. The user can remove video clips or replace the existing clips with recommended alternatives (Figure 3). Find Similar Alternatives is controlled by the options “by Characters”, “by Theme”, etc, as shown in Figure 4. By default, all the criteria are used. The user can replace existing clips in the timeline, or open the annotation details.

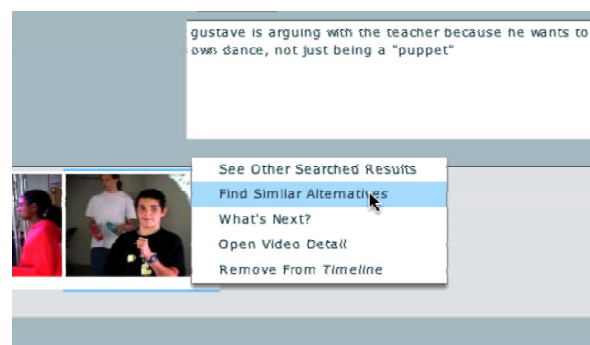


Figure 3: Finding Similar Alternatives for a clip

Existing stories can be extended by dragging video clips from the imported video list, typing story descriptions, or using the “What’s Next?” recommendation. Suppose the user chooses “similar emotions” as the only criterion for

the two. In that case, the best result would be “Louis wants to put ninjas in the background during the dance. Tiffany is not so enthusiastic about the idea.”, which does share similar emotions (“upset”: “not enthusiastic”). However, if the user selects “similar theme” as the only criterion, then the best search result would become “Glorianna asks Jacques about how he plans to involve the children in making decisions about different aspects of the performance.” which is, surprisingly, a nice transition from the previous clip “Gustave is upset because he wants to experiment with presenting the dances in different ways, but so far Jacques has been telling the kids what to do. Gustave just wants the chance to make something up himself.” This illustrates how Storyed Navigation's "What Next" feature can lead to the unexpected discovery of a meaningful and appropriate transition between scenes.

Annotating Video Sequences

Now, we return to considering the interface for annotating clips. Figure 5 shows the system’s annotation interface in the Advanced Mode. A simpler Basic Mode displays only the left half of this window.

Annotators can choose one of the videos from a scrolling list, watch it, and type a free-text story description for the video. Here, the description includes “Larissa talks to Seymour about difficulties...”. The system tries to identify the characters, actions, roles and emotions. In the Parsed Description pane, Larissa is identified as a Subject, Talk as an Action, and Seymour as an object. Character roles can be carried over from previously imported videos, where Seymour is identified as a Professor, and Larissa as a Graduate Student. Words that convey emotional content are particularly important for storytelling, and in this example, “difficulties” is identified.

The Other Sequence Information pane shows such optional information as theme, location, date, importance level (major vs. subsidiary), parsed or inferred from the story

description. The scene can be rated by the user as Major or Subsidiary in importance. The story Structural Role has eight different options: “begin”, “unfold”, “rise”, “transition”, “conflict”, “resolution”, “converge”, and “ending”. These choices originated from our personal experience in film editing. The user may edit any choices made by the system at any time.

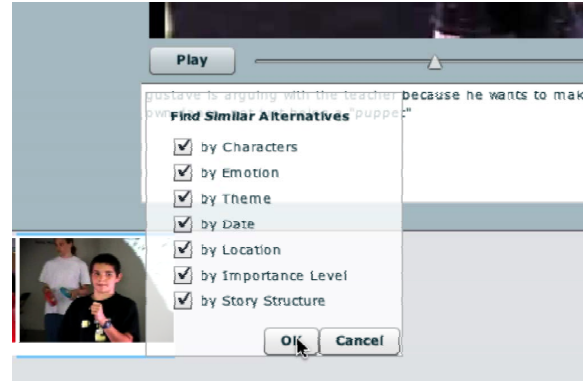


Figure 4: Finding Alternatives for an Existing Video

If the system is given the sentence, “Louis wants to put images of ninjas in the background during the dance. Tiffany is not so enthusiastic about the idea.”, the emotion of Tiffany is labeled as the most approximate terms that the system is able to find using our affect sensing algorithm. The system will fill in “sad, unhappy” for the emotion recognized in this sentence.

NATURAL LANGUAGE UNDERSTANDING IN STORIED NAVIGATION

Previously, we have described how users can annotate clips with natural language and story attributes, and how users can use natural language input to construct stories. This section details how the story understanding works, and how features are extracted.

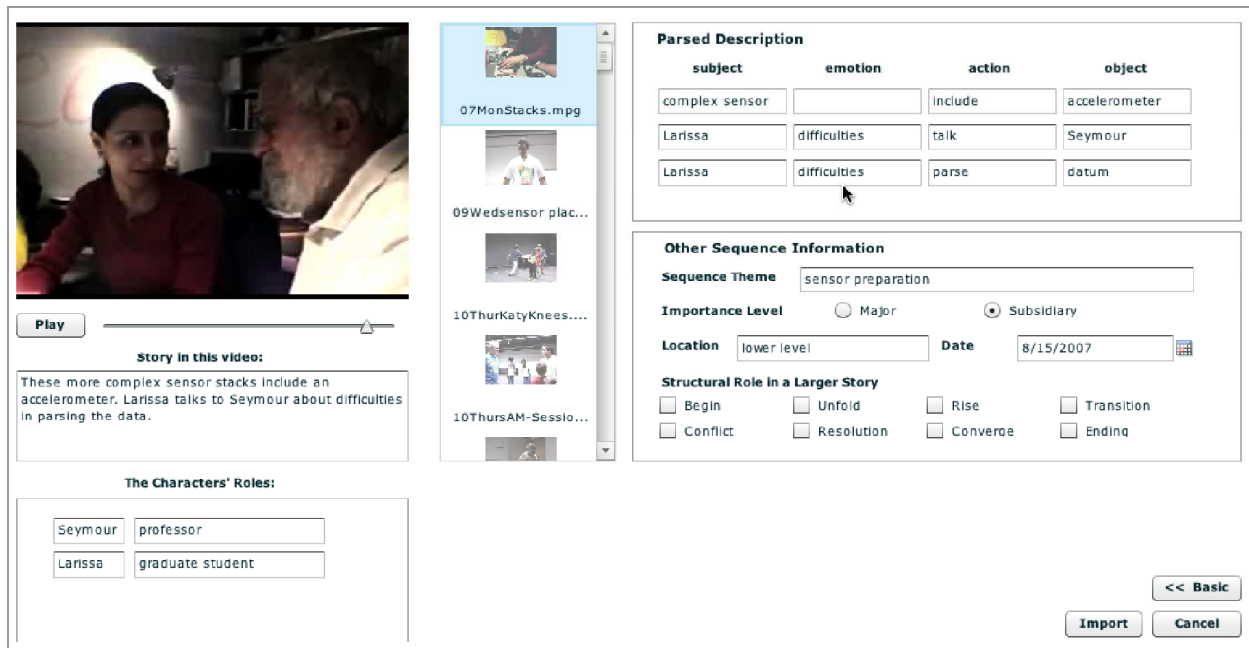


Figure 5: Annotation: Advanced Mode

By using commonsense reasoning techniques, from a free-text story we are able to extract the characters (semi-)automatically, the actions they perform, their emotions, and the theme. The story structural role of a sequence, and the social status of each character, still rely on users' manual specification. We fully realize that natural language processing (NLP) is, at present, not a perfectly reliable technology. First, we note that the system is "fail-soft", in that, if the natural language system does not properly extract relations or make suggestions, the user is still free to use the video composition facilities as in a more conventional video editor such as iMovie or Premiere. Second, in the annotation interface, the user has the opportunity to correct relations discovered by the system or add additional relations. Third, the contribution of this paper is in the interface that integrates an interactive video editor with novel use of NLP, not in NLP per se. We believe the results of our user studies, presented later, show that the current state of NLP is "good enough to get off the ground", in that it is not too distracting or frustrating to users, and we expect NLP itself to improve over time. We do present novel aspects of our use of NLP, including integration with Commonsense reasoning and story structure.

Figure 6 shows the Story Parsing Algorithm for extracting features. The input sentence t is first chopped into several sub-sentences using MontyLingua, a natural language parser [16]. Then, for each of the sub-sentences, the "Basic-Parsing" function produces a subject character, the character's emotion, the action, and an object character. It also finds other potential human characters (e.g. pronouns, proper names, roles) that coexist in the sub-sentence. For example, in the sentence "According to Tom's suggestion, we decide to go to that restaurant", "Tom" is an additional character the system should be aware of, even though he is not the main character.

```

procedure StoryParsing(TextString  $t$ )
     $sentence\_info$  = InitiateSentenceAnnotation( $t$ )
     $sub\_sentences$  = ChopIntoSubsentences( $t$ )
     $primitive\_subsentence\_reps$  = new Array
    foreach  $s$  in  $sub\_sentences$  do
         $parsed\_result$  = BasicParsing( $s$ )
         $primitive\_subsentence\_reps.append(parsed\_result)$ 
    foreach  $p$  in  $primitive\_subsentence\_reps$  do
        for  $character$  in ( $subject$  of  $p$ ,  $object$  of  $p$ ) do
             $gender$  = RecognizeGender( $character$ )
             $character$  = FindNameForPronoun( $character$ )
             $character$  = ReplOwnershipWOrigCharacter( $character$ )
             $sentence\_info.characters.append((character, gender))$ 
             $sentence\_info.subsentence\_primitives.append(p)$ 
         $sentence\_info.concept\_rep$  = FindSentenceConceptRep( $t$ )
    end procedure

```

Figure 6: Pseudo Code of the Story Parsing Algorithm

The emotion sensing algorithm is based on Liu et al.'s affect sensing algorithm [15]. Affect is described by a triple of numbers between +10 and -10, representing the PAD (Pleasure-Displeasure, Arousal-Nonarousal, Dominance-Submissiveness) model [17]. For example, the PAD vector of "love" is [8.72, 6.44, 7.11], and the PAD of "sad" is [1.61,

4.13, 3.45]). In Liu's approach, the PAD vectors for a set of keywords are used as the "emotion grounds", and a spreading activation algorithm propagates these values through ConceptNet with a decaying weight d ($d = 0.25$ in our system). A PAD value for a given node is computed by averaging all the incoming vectors. Negated expressions (e.g. "is not") are handled by computing a complementary vector [10.0-p, 10.0-a, 10.0-d]. For example, for the sentence "Tiffany is not so enthusiastic", the system will find "sad, unhappy" as the new emotional terms by using the original term "enthusiastic".

Three functions are employed to process each of these sub-sentences' subject and object characters: "Recognize-Gender", "Find-Name-For-Pronoun", and "Replace-Ownership-With-Original-Character". First, the subject and object characters are classified as one of the five "genders" using WordNet: "male", "female", "group of people", (e.g. "team", "friends", "My father and Joe"), "uncertain gender" (e.g., "Alex", "driver"), and "inanimate object". Then, for pronoun characters, the system will replace them with existing subject or object characters that share identical genders as properly as possible. For the example, in the sentence "Bob and Katy are my colleagues. They are both talented people.", "They" will be changed to "Bob and Katy". Finally, for an example sentence "I met Janet today. Her skirt was gorgeous.", the function system modifies "Her skirt" into "Janet's skirt".

The system constructs a sentence concept representation for this sentence in "Find-Sentence-Concept-Representation". It starts by, for each word, deriving a set of synonyms from WordNet using part-of-speech information found by MontyLingua. These synonyms are filtered using the *conceptual relevance* with a constant threshold, and the remainder are considered consistent with the word's meaning in the sentence. The conceptual relevance between two terms is defined as the overlapped nodes of their neighborhood in ConceptNet within a distance r . We use this technique to filter out certain synonyms derived from WordNet, because some are used less frequently than others. For example, WordNet returns ["get, acquire", "pay", "get, acquire", "believe", "be"] as synonyms for the verb "buy", in which "believe" and "be" will be filtered out. Then, it finds all the related concepts in ConceptNet for every original term and its synonyms, and each of these concepts will be rated by 1) the relevance scores given by ConceptNet, times 2) the inverse of the "familiarity" scores, or the usage frequency, given by WordNet. This is because, less-frequently used words (e.g. "negotiate") are more distinctive than often-used ones ("go") in determining similarity. Finally, the result found for the noun "party" include "involve", "wedding", "pool party", "party", "social gathering", "have brew beverage", etc.

While the derived characters, emotions, etc. are directly used as the story features, the concept representation is only the default version of a story's theme. When a new video is imported with a piece of story annotation, the system takes its derived concept representation to find the conceptually closest video sequence existing in the system's corpus, and use the linear combination of its original concept representation and this closest video's theme as its new theme if their concept representations are similar enough. This is because, the theme of the existing sequence may be manually changed by the users through the annotation interface, so it may be more precise to use this combination

```

procedure VideoComposition (TextString t, ParsedResult p)
  result_video_string = new Array
  potential_cs = DeterminePotentialCharacters(p)
  candidate_sets = EnumCandidateSets (potential_cs)
  foreach cset in candidate_sets do
    foreach  $\varphi$  in p.primitive_subsentence_reps do
       $\varphi$  = ReplaceCharactersWithCandidates(cset)
      sequences = FindSequences( $\varphi$ )
      sorted_seqs = SortSequences(sequences,  $\varphi$ )
      result_videos = AddBestSeqToResults (sorted_seqs)
      subsidiary_results = GetSubsidiaryResults(sorted_seqs)
    foreach video in result_videos, subsidiary_results do
      video = BuildMatchedStoryDescription(video, t, p)
    return result_videos, subsidiary_results
end procedure

```

Figure 7: Pseudo Code of the Video Composition Algorithm

as its default theme before the user makes any adjustments. Eq. 1 illustrates the similarity metric between two representations. For each concept c that coexists in both concept representations CR_1 and CR_2 :

$$S_{CR} = \sum (v_{c,CR_1} \times v_{c,CR_2}) \quad (1)$$

where v_{c,CR_i} denotes the value of concept term c in the concept representation CR_i . Thus, only concept terms that coexist in both representations will be taken into account.

EDIT-BY-TYPING

We wanted to provide a way of using the system that achieves acceptable results with the least possible complexity for the user. Our "Edit-by-Typing" facility allows the user to simply type a story in English, and a sequence of video clips is automatically returned. It is achieved by the Video Composition Algorithm, shown in Figure 7. After the system runs the story parsing algorithm on the story, it will look for the "potential characters" using the extracted characters. That is, if a character is not specific, say, "boy", the "potential characters" for this character will be any existing character whose description can be understood as "boy"; Otherwise, it will be the character itself. Then, the system enumerates all the candidate sets for the characters. If "Louis", "Tiffany", "Henri" are the only "child" characters in all the existing videos and "Jacques", "Duffin" are the only "teachers", then for the input "the teacher is watching the child improvising." there will be six candidate sets including ["Jacques", "Louis"], ["Jacques", "Tiffany"]...etc.

After the candidate character sets are enumerated, the system performs the following functions for each sub-sentence, and for each candidate set: 1) "Replace-Characters-With-Candidates", 2) "Find-Sequences", 3) "Sort-Sequences", 4) "Add-Best-Sequence-To-Results", and 5) "Get-Subsidiary-Results". First, the system replaces all the characters with the candidates. Then, the system finds annotated video sequences that contain all these characters. If any emotional terms exist in the input story description, the system ignores the sequences that

have no emotional terms for the corresponding subject characters. Third, the system calculates the similarity between the input story and each of the found sequences, filtering out the ones whose similarity values are below a threshold, and sorting the remaining ones. The similarity metric is shown in Equation 2, where S_{CR} is the similarity between two theme representations, S_A is the similarity between two affect (PAD) vectors, T_{CR} and T_A are the respective lowest thresholds, and μ_{CR} and μ_A are the respective constant parameters.







$$\text{If } S_{CR} \geq T_{CR} \wedge S_A \geq T_A, \text{ then } S = \mu_{CR}S_{CR} + \mu_A S_A \\ \text{else } S = 0 \quad (2)$$

The affect difference between the two stories is a constant K subtracted by the geometric difference D of the two PAD vectors, shown in Equation 2-a.

$$S_A = K - D(A_1, A_2) \quad (2-a)$$

Then, the system checks whether the best video sequence in the sorted list exists in the result video sequence. If not, it will append this sequence to the video sequence. The rest of the videos found by the algorithm will be collected in to the "subsidiary result list" for potential later use. Finally, the

Table 3: Examples of Using the Edit-by-Typing Function

Input	Returned Sequences & Their Respective Annotations	
"gustave is arguing with the teacher because he wants to make his own dance, not just being a puppet"		1. Gustave is upset because he wants to experiment with presenting the dances in different ways, but so far Jacques has been telling the kids what to do. Gustave just wants the chance to make something up himself
		2. Director Henri would like to thank everyone involved in the creation of "RoBallet." Although the dance was inspired by Louis, Jacques reminds them that Gustave had the idea of making up their own moves first.
"Louis is another child. In this workshop, he enjoyed how to dance, make animations, and even film other people using a video camera."		1. Louis asks Jacques if they can put their improvised dance in the show.
		2. Louis is making progress with his circle animation . Here, he explains his colorful pattern, and how simple it is to create .
		3. Louis films Seymour, who is having trouble with one of the laptops.
		4. Lilly surrenders her camera to Louis , who runs around filming the participants of the workshop. In this sequence entirely filmed by Louis , he experiments with zooming and unusual camera angles. Besides mastering the art of dance and the science of programming, Louis is a budding filmmaker.

system finds textual matches between the input story and that of each of the selected videos, in order to show how they are matched in the interface.

The interface of the edit-by-typing function is shown in Figure 2. Table 3 gives two examples of free-text story inputs and their respective responses. The first input produces two video sequences, whereas the second produces four. The bold-type terms are the matched terms shown on the interface. These sequences are the best results retrieved by the system, and are placed in the timeline automatically. For more results, the user simply needs to click on any of the sequences in the timeline and choose the “More Searched Results” option, and all will be listed on the right of the interface.

EDIT-BY-RECOMMENDATION

There are two types of recommendation functionalities, namely, “Find Similar Alternatives” and “What’s Next?”. The former is used to replace the currently selected sequence, whereas the latter finds sequences that can be juxtaposed after the selected one. The story features are used as the recommendation criteria, as shown in Figure 4. For finding similar alternatives, the system performs a high-level search using the selected criteria, e.g. a character search or emotion search. The only exception is that the theme search is incorporated into character search, to assure that similar emotional states are happening to the identical subject characters.

Table 4-a: Target sequence for “What’s Next?” Recommendation





	Gustave is upset because he wants to experiment with presenting the dances in different ways, but so far Jacques has been telling the kids what to do. Gustave just wants the chance to make something up himself
--	---

Table 4-b: Selected Criteria and the Respective Results for “What’s Next?” Recommendation

Criteria	Returned Sequence and Annotation	
Similar Character		Jacques asks Gustave to look directly into the camera as he dances so the audience can get a good look at his face.
Similar Emotion		Louis wants to put images of ninjas in the background during the dance. Tiffany is not so enthusiastic about the idea.
Similar Theme		Glorianna asks Jacques about how he plans to involve the children in making decisions about different aspects of the performance.

The “What’s Next?” recommendation is similar to “Find Similar Alternatives”, except for the two criteria, namely, “date” and “story structural role”. Referring to the interface shown in Figure 1, the option labels for these two criteria are “following dates” and “continued story structure”, as opposed to the others used to find videos of similar features. To look for video sequences that have “following dates”, the system simply finds sequences whose dates are after the selected one. Whereas for finding “continued story structure”, the system performs story structure search using a new 8-ary vector, each of whose binary numbers is shifted one-step forward from the selected video’s original vector. For example, if the selected video has a story structural role

(1,0,0,0,1,0,0,0), or (“begin”, “conflict” and none of the others), the system will perform story structure search using the new vector (0,1,0,0,0,1,0,0), or (“unfold”, “resolution”).

Table 4 shows three “What’s Next?” results for an example target sequence (Table 4-a), based on three different criteria (Table 4-b). First, the user selects “Similar Characters”, and the system returns a sequence that has both “Jacques” and “Gustave”, identical with the target one. In row 2, or the second case, Tiffany’s “not so enthusiastic” emotion state can correspond to Gustave’s being “upset”, which may be useful for building an intense atmosphere. Finally, if the user selects “similar theme” as the only criterion, then the result is the sequence in row 3, which is, surprisingly, a nice transition from the previous clip. From a filmmaker’s view point, we regard it an interesting, and potentially useful cut.

Besides the above functions, users can also search for videos using any of the story features. That is, calculating the number of coexisting characters, the emotion similarity between the input and the video subjects’ emotional states, etc, the system will find and sort videos accordingly.

OBSERVATION & EVALUATION

The video corpus used in our user study was a video collection of “Roballet”, a two-week workshop. There are 81 annotated video sequences in the database, ranging from tens of seconds to about three minutes.

Study 1: Making Video Stories

The goal of this study is to investigate whether the system helps users to develop their story threads. Two subjects participated this study, each with 50-100 hours video editing experience. The subjects received a tutorial, two warm-up examples, and two real storytelling tasks. After the study was completed, each subject was asked to fill a questionnaire of 19 questions on the 5-point Likert scale, and was then interviewed for about twenty minutes. The instructions for the storytelling tasks are:

Story 1-A, “*Learning in Roballet*”: *This corpus is about Roballet, a dance workshop where the researchers try to investigate new ways of learning. Learning can take place by creating things, by emulating others, by collaborating, or even by doing the opposite of what’s being told. Please make a story about the topic “Learning in Roballet”.*

Story 2-A, “*Technology and Art in Roballet*”: *Roballet is also an event where a new form of art is created by introducing technology into the creative process. Subthemes might be “Frustration with technology” or “The next, exhilarating era of art”.*

Subjects 1 and 2 participated in Roballet as a staff member and a graduate student, respectively. Both of them have 50-100 hours of experience of video editing. Their most familiar editing systems are iMovie and Adobe Premiere, respectively. Both subjects “strongly agree” they can use the system to edit videos, both find using the system “very enjoyable”. They both find the “edit-by-typing” function “very useful”, and both think “search by character”, “search by story structural role”, and “search by keyword” “useful” or “very useful”. They both think that the interface is “intuitive” or “very intuitive”, and neither finds any function “not useful” or “not useful at all”. The directions of questions asked during the interviews, including:

- i. *The Stories* (e.g. “Tell me about the stories that you made.” “What were the stories about?”)
- ii. *The Development of the Stories* (e.g. “How did you decide the characters of this story?” “How did you find the focus of this story?”)
- iii. *The Functions* (e.g. “What were the most useful functionalities in the process and why?”, “Which of the search functions did you find the most useful?”)
- iv. *Difference from Past Experiences* (e.g. “What is the major difference between this storytelling experience from your past experiences?”)

From the interview, we learned that Subject 1’s learning story was made by first typing, “Learning in Roballet, exploring body language along with mathematic language”, to which the system responded with a sequence where the staff talked about their own perspectives of how dance and technology should be emphasized without detouring from the main goal of the workshop. Then, by asking for “What’s Next?” based on the first result only, she found almost all the sequences she needed to complete the whole story. Here are two quotes from Subject 1 with regard to this result:

- i. *“This is right on the topic I asked. It’s a perfect clip”*
- ii. *“...the first one is so rich. It actually has every single component I put in my sentence. That’s the ‘What’s Next’ that goes back to explain it...each of the little segments is there. So I think it works nicely”*

On the other hand, according to Subject 2, the difference between the instructions is that the second one gives a hint of controversy. His Story 1-A was “not really a story, just like an informative video about learning”, as opposed to his comment on Story 1-B, “This one has a little bit more of a story, right?...Seymour was excited, the dancers are excited, everyone was excited, but then the technology doesn’t work! ...Why is the thing so slow? Things are not working the way they should...?”. He said “I was motivated. And I was fully aware of the difference”, emphasizing how the different directions influence the tasks. For functions used by Subject 2, keyword search was used the most for Story 1-A, whereas character search and story structural role were used a lot in Story 1-B. During the discussion on Story 1-B, he said “I sometimes just put in a word “programming”... like a Google search, because I have the best idea of the best clip, but it doesn’t pull out the one that resonated with it” and “The rise seem to be more accurate”, suggesting the possibility that the more “storied” functions may work better in a more storied situation.

From the interviews, many interesting observations were collected. Below we list three of the most interesting ones.

Improving Efficiency for Documentary Video Editing. The subjects pointed out that it would be much more difficult to use conventional editing software to build stories, because the time spent for understanding the material is much longer. To the question “What if you use iMovie from scratch?”, Subject 1 answered, “Oh, that would be incredibly hard... Because it’s...a two-week workshop. And they’re filming constantly... So now if I want to go with this topic, that means I have to watch, how many hours?... A lot.”, pointing out the efficiency that Storied Navigation confers.

New Ways of Organizing the Data. The subjects indicated that the ability to compose stories on the fly changed the

way they think about annotation and organizing their video collections: “...it’s hard to go back to something that has been done. But as you produce your material, and you know you can have it applied to a software like this, while you’re doing it you prepare it...at the end of the day you come, you just fit it to your software and you make your annotation quickly, and it is ready to use later.” which also raises an interesting question whether our system affects how people view annotation and manage their media files.

The Annotation Interface Helps Understanding the Logic Behind the Story. Another aspect about the annotation was also mentioned, “I also like the fact that I can look at the background, on how that clip came about, because once I know how the program is thinking, I can better adjust my questions.” The subject thinks that the annotation interface helps to phrase the input sentence in a more efficient way, which is an important finding as well.

Study 2: Browsing a Video Collection

Two hypotheses are to be verified: 1) “the system is helpful in the process of browsing an unknown video corpus” and 2) “when the subjects gain better understanding about how to use this system, they will browse the corpus in a more ‘storied’ way”. The seven subjects are all unfamiliar with the corpus but experienced in online video browsing. Again two pieces of instruction are presented, but the subjects are asked only to gain familiarity of the corpus by browsing videos using the system – similarly to using YouTube, not to create a new story. Each subject also participated in an interview session, and was asked to fill a questionnaire on a 5-point Likert scales.

All the seven subjects “agree” or “strongly agree” that they can use the system to browse videos, that the system helps them to find what they want, that they can use the system to edit videos – even though they were not asked to edit videos in the task. They say the interface design is “intuitive” or “very intuitive”. Five out of the seven subjects think using the system is “enjoyable” or “very enjoyable”, and all the subjects who have video editing experience agreed that using this system is “easier” or “much easier” compared to using other editing systems.

Table 5 shows the frequency of usage of the functions available to the users. The “Reposition” function stands for the drag-and-drop action that the users performed to change the sequences’ order in the timeline. The second column indicates the number of subjects using the respective functions, and the third stands for the total usage times of the functions. The underlined functions are those that were frequently used by most subjects, and only the Search by Date function went unused.

Edit-by-Typing is the most frequently used function. Nevertheless, different people used it in different ways. One subject said, “I started with the edit-by-typing function, but it’s hard for me to make long sentences...so then I just tried to look for things that I wanted from the result it gave me.” Another subject said, “I searched the related keywords: sensor, animation, dance...but then it’s too slow to do it one at a time, so I decided to type a whole sentence to make a sequence...” They responded that Edit-by-Typing gives flexibility for describing complex meanings that keyword search cannot accommodate.

Table 5: Functions Used in Study 2

Function Type	# Subjects Ever Used it	# Total Usages
Edit by Typing	6	52
Search by Theme	5	7
Search by Keyword	7	34
Search by Character	2	2
Search by Emotion	3	8
Search by Location	1	1
Search by Date (not used)	0	0
Search by Story Structural Role	3	10
Find Similar Alternative	4	7
Find What's Next	3	13
Reposition	5	44

Repositioning of the video sequences in the timeline is interesting because, the subjects were asked to find unfamiliar but interesting sequences and were encouraged to put these sequences in the timeline, but not asked to order them at all. However, it became one of the most popular functions. This suggests the possibility that the subjects have a sense of story flow even during the process of simply “finding” videos that interest them. One subject said, “Oh, I actually wanted to make a story... But why? It’s kinda weird... Maybe it’s just a habit I guess, when you write something you take care of the continuity.” The word *write* that he used indicated that he was already thinking like an author, even though the task was only browsing. We view it a significant change of perspective, which we doubt will happen using conventional browsing tools.

While the subjects regarded the Storied Navigation system as a (perhaps unconventional) search tool, they were quite attuned to the differences between it and keyword search used in Google or YouTube:

- i. The system helps people narrow down their rough ideas of manipulating the stories.
- ii. It allows users to organize selected clips directly in the search interface; they don’t need to jump back and forth between two interfaces.
- iii. The different forms of annotation clarify “how other people look at it” as well as how the system reasons about these sequences, and, accordingly, how users should modify their queries.
- iv. The experience is fun because it is a collection activity. As one subject said “It’s like online shopping. You can put something you want in the shopping cart”.
- v. It can be a new way of video blogging, because the experience of searching and arranging sequences is “not like introducing a restaurant to a friend. But it’s like you digest it a bit, and then present it to them”
- vi. “What’s Next?” is very helpful, as one subject said “YouTube really should add this in their website.”

DISCUSSION

The Accompanied Editor. One way to summarize our users’ experience is that they were able to compose stories or browse the corpus as if they were accompanied by an

experienced film editor as their guide. Though the system did not force them into any particular story path, it kept them focused on the question of “What’s next?” according to semantically meaningful story paths that captured and held their interest. It points to a future where media elements will be “fabricated” into stories for users to view, share, and relive their memories from, not just as fragmented pieces for passive viewing as they are today.

The Democratic Storytelling. Another way in which Storied Navigation could evolve is to serve as a way for viewers to consume a corpus directly, by dynamically composing their own story paths rather than relying on a path created by an external author. Users could traverse a corpus many times with different viewpoints. These kinds of systems can democratize the process of story creation, making everyone into a story author. Even if, from a professional’s perspective, each story is not so polished, a viewer may derive great pleasure from the process of exploration, which traditional cinema may not offer.

RELATED WORK

Many systems, from conventional timeline-based video editors such as Final Cut or Premiere, to form-filling template-based scripting systems such as Dramatica, have assisted users in composing video sequences. In previous work, however, either no path recommendation was provided [28], navigation in general story corpora was not supported [29], the mechanisms only performed keyword matching and no other understanding capabilities [8], or the users could only develop their stories using a small set of pre-designed paths [3].

Mindful Documentary [4] helps users in the real time video shooting process, rather than editing. Dexter and ConTour [8], Textable Movie [27], Vista [11], etc., do assist storytelling activities, but none of them uses commonsense reasoning or provides for story continuation. In the computer graphics community, the “schematic storyboards” of Goldman *et al.* [9], is a visualization technique that composes clips in a static image with arrows, outlines, and other visual annotation tools. Assa *et al.* [2] also tried to produce synopsis of the motion of 3D characters in still images, but neither of them focus on storylines or alternative narratives.

Computational analysis of stories or narratives uses terminology like “syuzhet” and “fabula” [20]. Cavazza and Pizzi’s also made a comparison between several narrative theories [6]. In automated video editing or summarization, Truong and Venkatesh presented a systematic review of existing methods of video abstraction, video summarization, and video skimming [25]. Few of these techniques can be used to generate stories that express causality or the transition of characters’ emotions. Recent works try to come up with novel representations such as camera motion, or a narrative structure graph [13], but they do not facilitate the process of collection-based storytelling using free-text, or recommendations based on the “ways to think”.

In terms of media annotation interfaces, Ossenbruggen and Hardman [19] proposed annotating temporal information by drawing lines, Appan *et al.* [1] presented a system that helps users to annotate videos stories by “who, what, when, where”. A direct predecessor to Storied Navigation is our own ARIA [14], which facilitates story-oriented annotation and retrieval from a still photo library.

Finally, some of the AI techniques we rely on have a long history. The commonsense reasoning tools that we use are ConceptNet [16], Monty-Lingua, and WordNet [10]. ConceptNet is a semantic network based on Open Mind Common Sense [23], a website that collects novice users' commonsense knowledge in English sentences. It is composed of 300,000 concepts and 1.6 million edges of 20 link types (e.g. "Effect Of", "Location Of"), and has functions for affect sensing and analogy [10, 15]. Our theory of "ways to think", based on our own experience in filmmaking, is closer to Schank [22], which uses indexing of stories, and the relationship between intelligent problem solving, memory, and storytelling.

CONCLUSION & FUTURE WORK

This paper presents *Storied Navigation*, an interactive video composition system that assists users in developing storylines in a brainstorming-like fashion. Using commonsense reasoning technology, three functionalities are implemented based on the "ways to think" in people's storytelling activities: edit-by-typing, edit-by-recommendation, and high-level search. A preliminary study conducted with seven subjects and an 81-video-clip corpus was encouraging, and we are planning to test the system with more subjects using materials of more variety.

In order to bring storytellers easier ways to understand their materials, and to come up with sequence orderings more coherent with the stories they really wish to tell, future work may be pushing it to the next level in terms of computational capability. For example, by finding ways to implement more "ways to think", or by creating plan recognition models that can infer what the users think about during the storytelling process – the story structure, in particular. To many of us, editing movies is enticing, but there are always times when we need a few hints. While we can't have perfect solutions, systems like *Storied Navigation* can help replace exasperation with inspiration.

Reference

- APPAN, P., SUNARAM, H. AND BIRCHFIELD, D. Communicating everyday experiences. *Proc of the 1st ACM workshop on Story Representation, Mechanism and Context* (2004).
- ASSA, J., CASPI, Y., AND COHEN-OR, D. Action synopsis: Pose selection and illustration. *Proc. SIGGRAPH 2005*, 24, 3, 667–676.
- BANGSØ, O., JENSEN, O. G., JENSEN, F. V., ANDERSEN, P. B., AND KOCKA, T. Non-linear interactive storytelling using object-oriented Bayesian networks. *Proc of CGAIDE 2004*.
- BARRY, B. *Mindful Documentary*. PhD Thesis, MIT Media Lab. 2005.
- BOCCONI, S., NACK, F., AND HARDMAN, L. Vox Populi: a tool for automatically generating video documentaries. *Proc. of ACM HT 2005*.
- CAVAZZA M. AND PIZZI, D. Narratology for Interactive Storytelling: a Critical Introduction. *Proc of TIDSE 2006*.
- DAVENPORT, G., BARRY, B., KELLIHER, A., AND NEMIROVSKY P. Media Fabric: A process-oriented approach to media creation and exchange. *BT Technology Journal* (2004) 22, 4.
- DAVENPORT, G. AND MURTAUGH, M. Automatist storyteller systems and the shifting sands of story. *IBM Systems Journal* (1997), 36, 3.
- GOLDMAN, D. B., CURLESS, B., SALESIN, D., AND SEITZ, S. M. Schematic Storyboarding for Video Visualization and Editing. *Proc of SIGGRAPH 2006*, 25, 3, 862-871.
- FELLBAUM, C. (Ed). *WordNet: An Electronic Lexical Database*. MIT Press. 1998.
- FIGA, E. AND TARAU, P. The VISTA Project: An Agent Architecture for Virtual Interactive Storytelling. *Proc of TIDSE 2003*.
- GOGUEN, J. AND HARRELL, F. Foundations for Active Multimedia Narrative: Semiotic spaces and structural blending. In *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*. 2004.
- JUNG, B., SONG, J. AND LEE, Y. A Narrative Based Abstraction Framework for Story-Oriented Video. *ACM Transactions on Multimedia Computing, Communications, and Applications 2007*, 3, 2.
- LIEBERMAN, H. AND LIU, H. Adaptive linking between text and photos using common sense reasoning. *Proc of AH 2002*.
- LIU, H., LIEBERMAN, H., AND SELKER, T. A Model of Textual Affect Sensing using Real-World Knowledge. *Proc of IUI 2003*.
- LIU, H. AND SINGH, P. ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal*, (2004) 22, 4, 211-226.
- MEHRABIAN, A. *Manual for a comprehensive system of measures of emotional states: The PAD Model*. Available from Albert Mehrabian, 1130 Alta Mesa Road, Monterey, CA, USA 93940. 1998.
- MINSKY, M. *The Society of Mind*. Simon & Schuster. 1988.
- OSSENBRUGGEN, J. v. AND HARDMAN, L. Semantic Timeline Interfaces for Annotated Multimedia Assets. *Proc. of EWIMT 2005*.
- PROPP, V. *Morphology of the Folktale*. Texas Press. 1968.
- ROSSI, F. <http://www.documentarydoctor.com>
- SCHANK, R. *Tell Me a Story: A New Look at Real and Artificial Intelligence*. 1991.
- SINGH, P., LIN, T., MUELLER, E. T., LIM, G., PERKINS, T., AND ZHU, W. L. Open Mind Common Sense: Knowledge acquisition from the general public. *Proc. of ODBASE 2002*.
- TONG, X., LIU, Q., DUAN, L., LU, H., XU, C., AND TIAN, Q. A Unified Framework for Semantic Shot Representation of Sports Video. *Proc of ACM MIR 2005*. 127-134.
- TRUONG, B. T. AND VENKATESH, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP 2007)*, 3, 1.
- TUFFIELD, M. M., MILLARD, D. E., AND SHADBOLT, N. R. Ontological Approaches to Modeling Narrative. *Proc. of AKT DTA Symposium 2006*.
- VAUCELLE, C. AND DAVENPORT, G. An open-ended tool to compose movies for cross-cultural digital storytelling: Textable Movie. *Proc. of ICHIM 2004*.
- BROOKS, K. M. *Agent Stories: Authoring Computational Cinematic Stories*. PhD Thesis, MIT Media Lab. 1999.
- MATEAS, M. and STERN, A. Natural Language Understanding in Façade: Surface-text Processing. *Proc of TIDSE*, Darmstadt, Germany, June 2004