Corpus Annotation for Narrative Generation Research

A Wish List

Pablo Gervás
Universidad Complutense de Madrid
Madrid, Spain
pgervas@sip.ucm.es

ABSTRACT

This paper outlines the type of annotation of narrative that would be of interest for the purpose of research into narrative generation. This is done by first charting some of the goals of narrative generation research, and then trying to identify the features that might need to be annotated to address those goals.

1. INTRODUCTION

The last decade of NLP research has seen a rise in importance of annotated corpora as research tools. Their applicability extends to providing better understanding of specific phenomena, as evaluation material for natural language processing (NLP) solutions, and as training material for machine learning approaches. In parallel, there has been a rise in interest in narrative as a subject for NLP research, in terms of analysis [5, 21], generation [2, 13, 27, 12] and evaluation [24]. In some cases, this has involved joint work between narratology and computer science departments [14].

It was only to be expected that this situation gave rise to efforts at producing annotated corpora for narrative. In recent times there have been a number of initiatives directed at generating corpora of narrative texts annotated in some form: with descriptive referring expressions [15], emotional information [10, 9], semantical information [6], structural, functional, and emotional aspects connecting discourse segments in a coherent story [16] and an integration of linguistic markup and semantic models of folk narratives [18]. In parallel, there have also been initiatives for the development of specific annotation tools [7], and the application of new methods to narrative corpora, such as the use of analogical story merging for deriving narrative morphologies [8], or the use of latent semantic mapping to organize a fairy tale corpus and apply recommendation algorithms to it [19].

The existence of such a broad range of possible annotations for narrative is indicative of a great potential in terms of possible applications of the resulting corpora, but also of a

significant risk. Unless some effort is made to coordinate in some way the various initiatives, the result in a few years time is likely to be a large set of mismatched corpora, each providing a different type of annotation for a different subset of texts. It also brings to light a different problem: each type of annotation will be geared to addressing a particular goal with respect to narrative. Two of the current initiatives ([7] and [18]) involve efforts to integrate different kinds of annotation, which is a positive contribution to the field. As narrative involves such a broad range of phenomena (extending well beyond all those already under research for natural language in general), it is also probable that, to be feasible, particular efforts will need to focus on subsets of the possible annotations.

This paper outlines the type of annotation of narrative that would be of interest for the purpose of research into narrative generation. This is done by first charting some of the goals of narrative generation research, and then trying to identify the features that might need to be annotated to address those goals.

2. RELATED WORK

Since annotated corpora have become established as an emerging tool of recognised importance, there have been numerous efforts to annotate text, both automatically, using NLP tools, and manually. The range of annotation now being carried out extensively covers a large number of phenomena concerning text in general (meaning disambiguation, coreference resolution, semantic roles, temporal expressions, spatial expressions, named entities...). A review of all this work is well beyond the scope of the present paper, but this should not be taken to imply that such work is irrelevant to the annotation of narrative. In fact, the annotation of narrative should begin with as thorough an annotation of the generic linguistic features of the corresponding text as possible. The rest of this section will focus on some concepts of narrative theory that are relevant to the annotation issue, and a brief review of relevant research on narrative generation which will inform the goals that the desired annotation might address.

2.1 Relevant Concepts of Narrative Theory

As Callaway [2] points out "current narrative theories are incapable of serving as the foundation for a comprehensive computational model that informs a decision algorithm for narrative generation". Nevertheless, it is important to consider some concepts of narrative theory that may help to

	English	French	Russian
what	story	histoire	fabula
how	discourse	discours	sjuzet

Table 1: Story and Discourse

stake out the problem. According to many theorists, narrative has two components: what is told (what narrative is: its content, consisting of events, actions, time and location), and the way it is told (how the narrative is told: arrangement, emphasis / de-emphasis, magnification / diminution, of any of the elements of the content). These have been named differently by different researchers, as described in Table 1. There are alternative analyses that postulate different subdivisions. Even between theories that agree on having just two levels of analysis there seem to be many subtleties that cast doubt on whether the same thing is meant by the different words. For instance some authors distinguish between the set of facts that characterise the situation being described (irrespective of whether they are actually mentioned in the discourse) and the set of facts that are mentioned in the discourse. There are also others who distinguish the linear sequence of facts that constitute the discourse and the actual text used to convey these facts linguistically. This lack of agreement on basic terminology presents a serious obstacle for researchers from the computational field trying to address the treatment of stories in any form.

An important number of the narrative-specific aspects of discourse were identified by Genette [11].

Narrative distance can involve *narratised speech* ("He confided in his friend, telling him about his mother's death."), *transposed speech, indirect style* ("He confided to his friend that his mother had passed away."), *transposed speech, free indirect style* ("He confided to his friend: his mother had passed away."), *reported speech* ("He confided to his friend: "My mother passed away."").

The narrator can play different functions when conveying text: *narrative function* (he just tells), *directing function* (he interrupts the story to comment on its organization), *communication function* (he addresses the text's potential reader in order to establish or maintain contact with him or her), *testimonial function* (he comments on the truth, precision, or sources of the story, or his emotional involvement with it), and *ideological* function (he interrupts his story to introduce instructive comments or general wisdom concerning it).

Narrative distance and function of the narrator conform what is known as narrative mood.

Narrative voice captures whether the narrator is present or absent from the story he tells (and whether or not he is the hero of the story).

The time of narration captures the relation between the time of telling and the time being told about. It covers four kinds: *subsequent narration* (the narrator tells what happened in some past time), *prior narration* (the narrator tells what is going to happen at some future time), *simultaneous narration* (the narrator tells his/her story at the very moment it occurs) and *interpolated narration* (combines prior and simultaneous narration).

Narrative perspective or focalization is the way in which a narrator restricts what he is telling about a particular scene to what might have been perceived by someone present in that scene. There are three kinds of focalization: *zero focalization* (the narrator knows more than the characters), *internal focalization* (the narrator knows as much as the focal character) and *external focalization* (the narrator knows less than the characters).

Narrative voice, the time of the narration and narrative perspective conform narrative instance.

Narrative occurs at more than one level whenever a character in a story starts telling a story. Such nestings of stories within stories determine narrative levels.

Narrative time is characterised in terms of order, speed, and frequency of events.

Order is the relation between the sequencing of events as they actually occurred and their arrangement in the narrative. Any departure from the original chronological order is called anacrony. There are two types of anacrony: *analepsis* (the narrator recounts after the fact an event that took place earlier than the present point in the main story) and *prolepsis* (the narrator anticipates events that will occur after the present point in the main story).

In theatrical representations, the story is told at the same pace as it is happening (on stage). Speed involves introducing differences between the time the story takes to happen and the time taken to tell it. It is described in terms of four narrative movements: *pause* (the event-story is interrupted to make room exclusively for narratorial discourse such as static descriptions), *scene* (narrative time corresponds to the story's time, as in dialogue), *summary* (some part of the event-story is summarized in the narrative, creating an acceleration), and *ellipsis* (the narrative says absolutely nothing about some part of the event-story).

Frequency of events establishes the ratio between the number of times an event happens in the story and the number of times it is mentioned in the narrative. Events may be mentioned more times than they actually happened, or events that happened several times may be be told just once.

These parameters evolve over the course of a narrative, so that different spans of the narrative may have different values for each of these parameters.

2.2 Research on Narrative Generation
Research on narrative generation has a slightly different focus than narrative understanding research or language generation research. As this may impact on the type of annotation that would be most useful, a few relevant ideas are described here.

Narrative generation is an instance of natural language gen-

eration (NLG). As such, it inherits a number of generic issues from this field. Of these, the most significant is the subdivision into tasks [26] of *content determination* (deciding what to say), *discourse planning* (deciding on the order of presentation), *sentence planning* (deciding how to structure each sentence in its context of appearance) and *surface realization* (constructing the surface form text). However, NLG has in the past concentrated mostly on the production of non-narrative texts such as weather reports or instructive manuals. Nonetheless, there have been a number of efforts to develop generators specifically aimed at narrative.

STORYBOOK [2, 3] produced multi-page stories in the Little Red Riding Hood domain. It relied on elaborate natural language generation tasks such as narrative planning, sentence planning, discourse history, lexical choice, revision of drafts, a full-scale lexicon, and the FUF/SURGE surface realiser. Callaway introduces specific terminology for distinguishing what is told (the *fabula*: the sum of the factual content that constitutes the story) and how it is told (the *narrative stream*: the presentational ordering of the text). The differences and the relationships between them are clearly delimited. The narrative stream represents the information given explicitly. The fabula embodies the implicit information that underlies the narrative. Events in the fabula may be recoverable to a different extent from the narrative stream depending on how it has been generated. The fabula is a parallel knowledge structure, where information can be accessed in any order. The narrative stream is linear.

In his system Callaway models Narratorial Point of View and Narrator Modes, which are relevant to the narratological parameters mentioned in section 2.1. He also explains that the system requires a store of background cultural knowledge, including experience with previous stories told by others, ideas about how characters should act or talk, and fundamental knowledge of the world and how objects within it interact. Examples of this include simple facts like "trees have green leaves," "Grandma's house sits on the ground," and "a door serves as a portal between the inside and outside," even though they are never explicitly mentioned in any narrative. This he terms the story ontology.

The overall system involves a narrative planner (which actually generates a fabula and a narrative stream for the story), a narrative organizer (which produces a stream of sentential specifications) and a surface realizer (which converts these into text).

Callaway focuses on stylistics and mechanical realization during the production of text for an input story rather than the generation of fabula or narrative stream. The narrative planner is therefore simply specified and assumed to exist, with the actual implemented system dealing with the fabula and narrative stream as given inputs. Under this approach, a realtime narrative prose generator requires an algorithm that can intelligently combine information found in the narrative stream and fabula and convert it along with stylistic directives into narrative prose.

Lönneker [20] proposes an architecture for advanced NLG systems that handle narratives. In her paper, she reviews

Callaway's STORYBOOK as a possible architecture for this task, and concludes that most of the narratological parameters that ought to be considered are in Callaway's proposal left for the narrative planner to decide, as they come already specified in the narrative stream that acts as input. She then proceeds to argue that the most important decisions of a narratologically enhanced system concern the document planner with its content determination and document structuring subtasks. She outlines four tasks that such a narrative planner might address: managing various forms of anacrony, ellipsis, introduction of more than one narrative level, and handling point of view and focalization. To illustrate this point she includes a more detailed analysis of how narrative levels and narrative instance might be implemented.

Montfort's PhD thesis [22] developed the *nn* system for interactive fiction, which was designed to address the issue of narrative variation and therefore covered some of the narratological parameters mentioned in section 2.1. In an interactive fiction system the user controls the main character of a story by introducing simple descriptions of what it should do, and the system responds with descriptions of the outcomes of the character's actions. Within *nn*, the Narrator module provides storytelling functionality, so that the user can ask to be "told" the story of the interaction so far. The Narrator module of *nn* addresses important issues in storytelling that had not been addressed by previous systems: order of presentation in narrative and focalization. Instead of telling events always in chronological order, the *nn* Narrator allows various alternative possibilities: flashbacks, flashforwards, interleaving of events from two different time periods, telling events back to front... It also captures appropriate treatment of tense depending on the relative order of events being told to speech time, reference time, and event time. Focalization is handled by the use of different *focalizer worlds* within the system. Aside from the actual world of the interactive fiction system, *nn* maintains additional separate worlds representing the individual perspectives and beliefs of different characters. These can be used to achieve correct treatment of focalization (telling the story from the point of view of specific characters).

Montfort has developed a new version of his system, now known as *Curveship* [23], which is intended for people to use as an interactive fiction system (as opposed to the research prototype that *nn* was). Curveship can deal with distance, frequency, speed, as well as focalisation and order.

The works of Callaway and Montfort concentrate on the generation of narrative text from an already existing flow of events (the narrative stream received as input in the case of Callaway, and the game flow as dictated by the interactive fiction system in the case of Montfort). There are many story generation systems that focus on the generation of plots, relying on simple template-based solutions for the production of text renderings of these plots. Lönneker reviews some of these, discussing architectural options for a narratology-enhanced generator. The task of inventing a story plot (whether from scratch or to satisfy a given specification or set of input parameters) is slightly different from the task of generating narrative text for a given input flow of events. For a more extensive review of plot generation systems, the interested reader is invited to consult [12].

3. TOWARDS A CORPUS-BASED APPROACH TO NARRATIVE GENERATION

Corpus-based approaches have proved successful for many areas of NLP. Even natural language generation has seen its fill of statistical approaches based on training corpora, from early efforts based on syntactic annotation only, such as for instance [17], to more recent ones [1] which rely on multilevel annotation including semantic structures. However, the specific nature of narrative generation must be taken into account when devising such an approach.

3.1 Layers of Representation of a Story

The discussion presented in section 2.1 has shown that there is a lack of consensus on terminology in this field. However, the analysis of existing systems given in 2.2 allows us to identify a number of concepts that are relevant to a computational analysis of this problem. These concepts arise from differences in the nature of the computational operations and decision processes involved in deriving one from another, either during text understanding or text generation. Although literature on the subject is already overloaded with terminology, I will attempt to give tentative names to them to facilitate the subsequent discussion.

I will consider that the following possible representations can be associated with a given story:

text representation the linguistic realisation of the story

explicit representation the linear sequence of facts mentioned in the story (in some kind of conceptual representation)

underlying selected representation all facts relevant to the story that are mentioned in the explicit representation (the set of facts that are mentioned in the story, but not necessarily organised in a linear sequence and following a chronological partial order not necessarily equivalent to the one in which they appear in the story)

underlying extensive representation all possible facts relevant to the story (including causes, effects, emotional reactions, common knowledge, and generally all the additional material that will be inferred by a reader on reading the story)

The explicit representation corresponds roughly to Callaway's narrative stream. The underlying extensive representation would correspond to the set union of Callaway's fabula and his story ontology, and the underlying selected representation would correspond to the part of the fabula that gets mentioned in the narrative stream.

Some of these distinctions are particularly pertinent for our general goal, as automatic annotation using state of the art NLP tools involves uncovering some of these hidden layers from the given input, which is likely to be the original text of a story. Reviewing some of these in order of their likely application to the text, processes like syntactic parsing, semantic role labelling, named entity recognition, coreference resolution or word sense disambiguation would address part of the conversion from the text representation into the explicit representation. The identification of temporal expressions and temporal relations in a text may be seen as subtasks in the way towards identifying a chronological order for the facts in the story, independent from the order in which they appear in the story. This might produce the underlying selected representation. Finally, the automatic identification of the omitted information that a human reader might reasonably infer would correspond to producing the underlying extensive representation. This is a long sought goal of natural language understanding which seems yet a little far off.

It must be said that these representations are postulated strictly with the intention of addressing the problem from a computational point of view. No claim whatsoever is intended as to their cognitive plausibility. In fact, a large percentage of the problems that we will describe below arise from the fact that the human brain clearly operates in radically different ways. However, a computational analysis of the problem must handle such elements as we can represent and handle in symbolic terms.

Obviously all of these annotations will be useful for research on narrative, as they are common to narrative and text in general. This paper will focus on those aspects that are specific to narrative. These in general constitute aspects that address the relation between the different layers of a story.

3.2 Tasks in Narrative Generation

Narrative generation spans a number of tasks, from the creative activity of coming up with a new convincing story, to the craft of putting together a fluent text that conveys a given set of facts.

In the past, any program that produced the text for a story has been considered a story generator. This included programs that simply concatenated pre-written strings according to a story grammar, programs that used planning to build an underlying representation which was then converted into text using templates, or programs which took an explicit representation (much like STORYBOOK's narrative stream) and used heavy NLG to produce high quality text from it.

The introduction of these definitions allows a more finegrained classification. For instance, the simplest possible generators would be those that generate directly a text representation, with no conceptual representation equivalent to the explicit representation involved. Systems that generate directly a conceptual representation equivalent to an explicit representation (a conceptual representation that is already already linearised and ordered) would be less complex than those that generate an intermediate conceptual representation which is non linear and ordered differently from the final result). A detailed follow up of this line of work is beyond the scope of the present paper

Having an explicit sketch of the kind of representations that lie behind a story allows identification of some subtasks that might be involved in a complex generator:

invention production of an underlying extensive representation for a story

content determination production of an underlying selected representation from an underlying extensive representation

discourse planning production of an explicit representation from an underlying selected representation

telling production of a text representation from an explicit representation

The subtasks of content determination and discourse planning match those already identified for natural language generation. As in the case for NLG, it is not clear whether they can be carried out independently from one another. Decisions taken while carrying out one task may affect the other. This may indeed be true for all the subtasks that we are considering. Problems encountered during discourse planning or telling a story may be solved by additional processes of invention. This problem is similar to that of the subdivision of natural language generation into subtasks, and it need not be discussed here. In general terms, arguments equivalent to those presented by Reiter [25] could be put forward for and against such a subdivision as a useful abstraction. Additionally, invention may be applied at every layer of representation (inventing directly material for an underlying selected representation, an explicit representation, or even text).

Yet this set of tasks constitutes a useful set of tools for describing narrative generator. Most of the existing storytelling systems could be said to be carrying out invention, usually directly to an explicit representation, sometime to an underlying selected representation, and sometimes to an underlying extensive representation. But then very few of them rely on refined solutions for the content determination, discourse planning or telling subtasks. Important exceptions are STORYBOOK, which focuses exclusively on what we have termed the telling subtask, and nn/Curveship, which applies an interactive solution for the invention task, uses a quite refined solution for content determination and discourse planning (involving explicit treatment of several narrative parameters) and applies a very simple solution for telling.

There is an additional challenge for narrative generation that has not been explored in detail from a computational point of view. This is the craft of telling a set of facts as a story. With respect to our set of subtasks, it would correspond to taking an underlying extensive representation and applying to it processes of content determination, discourse planning, and telling (again, not necessarily in strict succession). This is in fact the kind of narrative generation that would be closest to traditional natural language generation.

However, a number of significant differences arise in this case.

Stories are traditionally understood as "sequences of events": a line of carefully alligned points in time. This is very rarely true. The underlying representation for most stories worth telling usually involves several events taking place simultaneously, in different locations, or even in the same location. If one were to assign a geometrical form to the underlying representation of a story it would hardly be a line. If you think of a story in terms of events, each event takes up a portion of time (duration) and each event can be associated with the portion of space from which it can be perceived (location). Even if you consider only a two dimensional representation for location, an event would have to be represented more as a volume than as a point. Additionally, the durations and/or the locations of different events may overlap. Once you consider a complete story, the geometrical representation for it would not even be a graph, but a set of heavily intersecting volumes of space/time.

The "linear story" mirage arises from the fact that stories are usually told as a sequence (of clauses, sentences, or even images if one considers films as stories). Language or film are by their very nature linear sequences, and stories wanting to be told by such means have to conform to this restriction. Traditional storytelling addresses this problem by drawing lines through the space/time volume that connect elements in a story. Each of these lines constitutes a narrative thread of the story. In this way, the volume is reduced to a graph. The next stage involves deciding how to traverse this graph as a sequence of linear paths through it, which result in the explicit representation of the story. With respect to our subtasks, the twin processes of drawing the threads and traversing the graph constitute instances of discourse planning. They probably do not happen sequentially but as a single complex interdependent operation.

Under this light, narrative parameters (and more specifically the conceptual operations that lead to the different formulations) become fundamental for understanding the storytelling task. For instance, focalization plays a central role in providing a rational way of partitioning the space/time volume into threads defined as what may have been perceived by a given focalizer. Different threads may be traversed by switching from one focalizer to another. This implies that in the underlying representation of a story events should carry additional information concerning their location, so that the focalization decisions can be informed. For correct focalization, decisions should be based not just on absolute concepts such as distance, but rather on whether an event can be perceived by a given focalizer. Many of the other phenomena covered by Gennette's narrative parameters can be seen as important tools for obtaining an appropriate linear sequence to convey the space/time volume of events of a given story. This is a key question for narrative generation research, and one that would clearly benefit enormous from the availability of annotated corpora.

3.3 Story Representations and Narrative Parameters

In the light of these definitions, many of the narrative parameters defined by Genette (described in section 2.1) constitute formal representations of specific relations between these different layers of representation of a story. Others bring into play additional material that we are forced to consider.

Narrative distance captures a number of choices open to the narrator in the way he decides to convey what has happened, when he makes the transition from the explicit representation to the text representation. Whereas natural language understanding is mainly concerned with problems of ambi-

guity (which of many possible representations for a sentence to assign as its interpretation), natural language generation faces problems of choice (which of many possible realizations for a given fact to employ in a certain context). The concept of narrative distance constitutes a very valuable source of information in this respect, and it could be very fruitfully used to inform consistent decisions in this respect across a given document.

Of all 5 functions of the narrator, only the narrative function is directly related to the actual story. This reminds us that narrative may include additional material that may need to be provide on the side for a generator and marked as such during annotation of a corpus. One should consider at what levels of representation such material would have to be included, and whether it should be marked as distinct from the actual story at all levels in which it appears. From the point of view of annotation or analysis, it may be tricky to discern in some cases the extent to which a certain element in the text concerns ideological material, for instance. In contrast, during generation such material would come from a different source to the rest of the content, so the difference is important.

Narrative instance captures the relation between the actual act of narration and the underlying story. Narrative voice will affect the use (in generation) and interpretation (in reading) of personal pronouns in particular spans. The time of the narration will affect the use of tense. These parameters affect the transition between the explicit representation and the text representation.

By changing focalization, however, different fragments of the underlying extensive representation may be selected to be included in the underlying selected representation. This can significantly alter the shape of the explicit representation and the text. It also plays a very significant role in the task of constructing a linear sequence to match a complex underlying story, as discussed above.

Narrative levels introduce the problem of recursion (a story within a story). This is a familiar problem in AI, but one must keep in mind that such recursion may involve a different four-layer representation of each level of embedding. Interactions between different levels may be problematic. This relates to the phenomenon Genette defines as *metalepsis* (characters from the main story appear in the secondary story).

The order parameter of narrative time captures the differences in partial ordering between the underlying selected representation and the explicit representation. This provides the flexibility required to traverse the graph implicit in an underlying selected representation (once threads have been established based on focalization) in a non-sequential way, and yet produce a linear explicit representation.

The speed parameter of narrative time captures the possibility that the underlying extensive representation and the underlying selected representation may differ in granularity. Each narrative move describes a different type of transition from one to another. In a scene, events in the underlying extensive representation get mapped to corresponding events in the underlying selected representation. In a pause, some of the static material (descriptive facts about elements that appear in the story) in the underlying extensive representation gets mapped to the underlying selected representation. In a summary, a set of events in the underlying extensive representation gets mapped to a much smaller set of events in the underlying selected representation. In an ellipsis, parts of the underlying extensive representation get omitted altogether.

Frequency of events captures the possibility that events appearing once in the underlying selected representation may be mentioned more than once in the explicit representation, or that events appearing several times in the underlying selected representation may be mentioned only once in the explicit representation.

3.4 A Corpus for Narrative Generation Research

This overview outlines some of the operations that need to be considered in a narrative generator in view of the selected stance on narrative theory. Read in a different way, it provides some clues as to what information might need to be annotated in a story for the resulting document to be useful from the point of view of narrative generation research.

The ideal case would be a story with the four layers of representation explicit and in which all the relevant narrative parameters fully annotated. As the parameters evolve in the course of a story, and they are determined by particular relations holding between different representation layers, some annotation scheme would have to be devised capable of capturing all this information.

From a corpus of such ideal cases, it would be possible to learn about:

- precise definitions of narrative parameters
- decision procedures for choices in transitions between representations (for generation)
- disambiguation procedures for interpretation (for understanding)

However, the construction of this ideal case would come up with several obstacles, most arising from the fact that only the text representation of a story is generally available.

First, it would be difficult to agree on an appropriate representation formalism for the conceptual knowledge involved in all layers other than the text representation. This is a long standing problem in AI and cognitive science that seems far from being solved.

Second, even if a conceptual formalism is chosen, the inferences involved in producing the underlying extensive representation would have to be carried out by human annotators. Because the range of possible inferences is large, a very low value of inter-annotator agreement is to be expected.

Third, the actual chronology of the underlying facts of a story is usually only vaguely specified in the text. This

problem has been addressed by Mani [21], who introduces *timelines* as an underspecified representation of temporal relations to deal with this problem. This solution allows for an economical representation that reduces disagreement, and it would help in annotating text. Exploiting such a solution to inform a narrative generator may need specific interfacing between whatever representation of time is being used in the input and these timelines.

Fourth, the theoretical definitions of the narrative parameters may not be amenable to precise definition in terms of relation between layers of representation. As the definitions were originally constructed without reference to such representations (which, as mentioned above, are particular to computational approaches to these problems), it is possible that any formal specification may fail to capture their original nuances[1].

Even considering all these obstacles, the task is not impossible. A simpler approach, based on annotating text with as much of this information as can be obtained would go a long way towards informing narrative generation research. The kind of information that might be obtained is quite extensive. As mentioned above, there are a number of NLP tools that could provide a very good starting point towards obtaining a good approximation to what I have called here the explicit representation of a story. Indeed, a large number of these are already integrated in the StoryWorkBench [7], an annotation tool specifically intended for narrative texts. Over this representation it would be interesting to add an additional layer of annotation to distinguish the specific functions of the narrator relevant to each span of the representation. This would help in sorting the specifically narrative material from all other contributions present in the text. As far as I know, this kind of annotation would have to be done manually. To obtain some form of underlying selected representation would require two additional steps: the identification of the temporal and spatial coordinates for the various elements (events but also descriptive material) identified in this explicit representation. Significant progress along these lines, and some interesting lines of future work with interesting potential, are reported in [21]. This kind of annotation would probably only be relevant for those spans of the material that correspond to a narrative function. If something like an explicit representation for a narrative text could be obtained, and annotated with additional information of narrator function, temporal information, and location information, it would in itself constitute a very powerful resource. This point has been forcefully argued by Mani in his recent book [21].

A possible extension would be to include additional annotations reflecting the values for all the remaining narrative parameters for the corresponding spans of text. This would provide a very valuable resource in itself for the study of narratology, possibly informing a more precise definition of these parameters in terms of the texts that they apply to. Such annotations would have to be carried out by hand, and the help of experts on narratology would be required. As mentioned above, a specific annotation scheme would have to be devised, capable of capturing the complex information

involved in each case. Once an initial corpus of this kind were available, it might be possible to write tools for automatic annotation, possibly with the application of machine learning techniques to such corpus used as training material.

Many of these narrative parameters in some way encode transition operations between the underlying representation of the story and the explicit representation. If both an approximation to the explicit representation and annotations for the relevant narrative parameters were available, they might constitute a very good starting point towards obtaining an approximation of the underlying representation of the story. By combining the temporal and spatial information in the explicit representation with the information on transitions encoded in the annotations of narrative parameters, it might be possible to obtain a more precise approximation to the underlying representation than would be obtained without the help of these additional annotations. This would directly be a very positive result for narrative understanding. The resulting approximation would probably need to be validated by hand, but once it were available it would constitute a very valuable source for research on narrative generation.

4. CONCLUSIONS

Narrative generation research would benefit significantly from the existence of corpora of narrative texts annotated with the information that is relevant to the field (in addition to annotations already in use for natural language processing). Much like in NLP, only a small subset of the information that would be desirable is likely to be available in the short term, due to the difficulties inherent in the annotation task. However, a number of narrative-specific extensions to existing annotations in NLP have been identified. These concern mostly the annotation of narrative texts with values for Genette's parameters for narrative discourse. Additionally, possible applications of some existing annotations (concerning temporal and spatial information) to narrative-specific issues have been described. These ideas are put forward as suggestions, as a wish list of what it might be useful for narrative generation researchers to find in annotated corpora of narrative. There are indeed many more aspects that would need to be covered for a detailed study of narrative. Issues such as emotion, author goals, intention, reader reaction, and figurative language have not been considered here, but they are probably as relevant to narrative as the more basic issues mentioned here. Nevertheless, even going as far as I did involved already a fair amount of speculation, so detailed discussion beyond the basics should be left till specific advances have been made. Such advances may support or cast doubt on some of the assertion made in this paper.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] B. Bohnet, L. Wanner, S. Mill, and A. Burga. Broad coverage multilingual deep sentence generation with a

---

[1]This may also be true of some of the interpretations presented in this paper.

stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

[2] C. B. Callaway. *Narrative Prose Generation*. PhD thesis, North Carolina State University, 1999.

[3] C. B. Callaway. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252, 2002.

[4] N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[5] N. Chambers and D. Jurafsky. A database of narrative schemas. In Calzolari et al. [4].

[6] D. K. Elson and K. R. McKeown. Building a bank of semantically encoded narratives. In Calzolari et al. [4].

[7] M. Finlayson. Collecting semantics in the wild: The story workbench. In *AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence*, pages 46–53. AAAI Press, Menlo Park, CA, 2008.

[8] M. Finlayson. Deriving narrative morphologies via analogical story merging. In *New Frontiers in Analogy Research*, pages 127–136. New Bulgarian University Press, Sofia, 2009.

[9] V. Francisco and P. Gervás. Ontology-supported automated mark up of affective information in texts. *Special Issue of Language Forum on Computational Treatment of Language*, 34(1):23 – 36, 2008.

[10] V. Francisco, R. Hervás, and P. Gervás. Two different approaches to automated mark up of emotions in text. In *AI-2006*, pages 101–114, Cambridge, England, December 2006 2006. Springer Verlag, Springer Verlag.

[11] G. Genette. *Narrative discourse : an essay in method*. Cornell University Press, 1980.

[12] P. Gervás. Computational approaches to storytelling and creativity. *AI Magazine*, 30(3):49–62, 2009.

[13] P. Gervás, B. Díaz-Agudo, F. Peinado, and R. Hervás. Story Plot Generation Based on CBR. *Knowledge-Based Systems. Special Issue: AI-2004*, 18:235–242, 2005.

[14] P. Gervás, B. Loenneker, J. C. Meister, and F. Peinado. Narrative models: Narratology meets artificial intelligence. In *International Conference on Language Resources and Evaluation. Satellite Workshop: Toward Computational Models of Literary Analysis*, pages 44–51, Genova, Italy, 2006.

[15] R. Hervas and M. Finlayson. The prevalence of descriptive referring expressions in news and narrative. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 49–54, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[16] O. Y. Kwong. Constructing an annotated story corpus: Some observations and issues. In Calzolari et al. [4].

[17] I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *ACL-36: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 704–710, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[18] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado. Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In Calzolari et al. [4].

[19] P. V. Lobo and D. M. de Matos. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In Calzolari et al. [4].

[20] B. Lönneker. Narratological knowledge for natural language generation. In G. Wilcock, K. Jokinen, C. Mellish, and E. Reiter, editors, *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005.

[21] I. Mani. *The Imagined Moment. Time, Narrative, and Computation*. University of Nebraska Press, Lincoln, Nebraska, 2010.

[22] N. Montfort. *Generating narrative variation in interactive fiction*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 2007.

[23] N. Montfort. Curveship: an interactive fiction system for interactive narrating. In *CALC '09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 55–62, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[24] F. Peinado and P. Gervás. Evaluation of automatic generation of basic stories. *New Generation Computing*, 24(3):289–302, 2006.

[25] E. Reiter. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Seventh International Workshop on Natural Language Generation*, pages 163–170, Kennebunkport, Maine, USA, 1994.

[26] E. Reiter and R. Dale. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, 1997.

[27] M. Theune, N. Slabbers, and F. Hielkema. The automatic generation of narratives. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. V. Eynde, editors, *Proceedings of the 17th Conference on Computational Linguistics in the Netherlands (CLIN-17)*, pages 131–146. LOT Occasional Series 7, LOT, Utrecht, 2007.