

# Learning Narrative Morphologies from Annotated Folktales

Mark Alan Finlayson

markaf@mit.edu

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

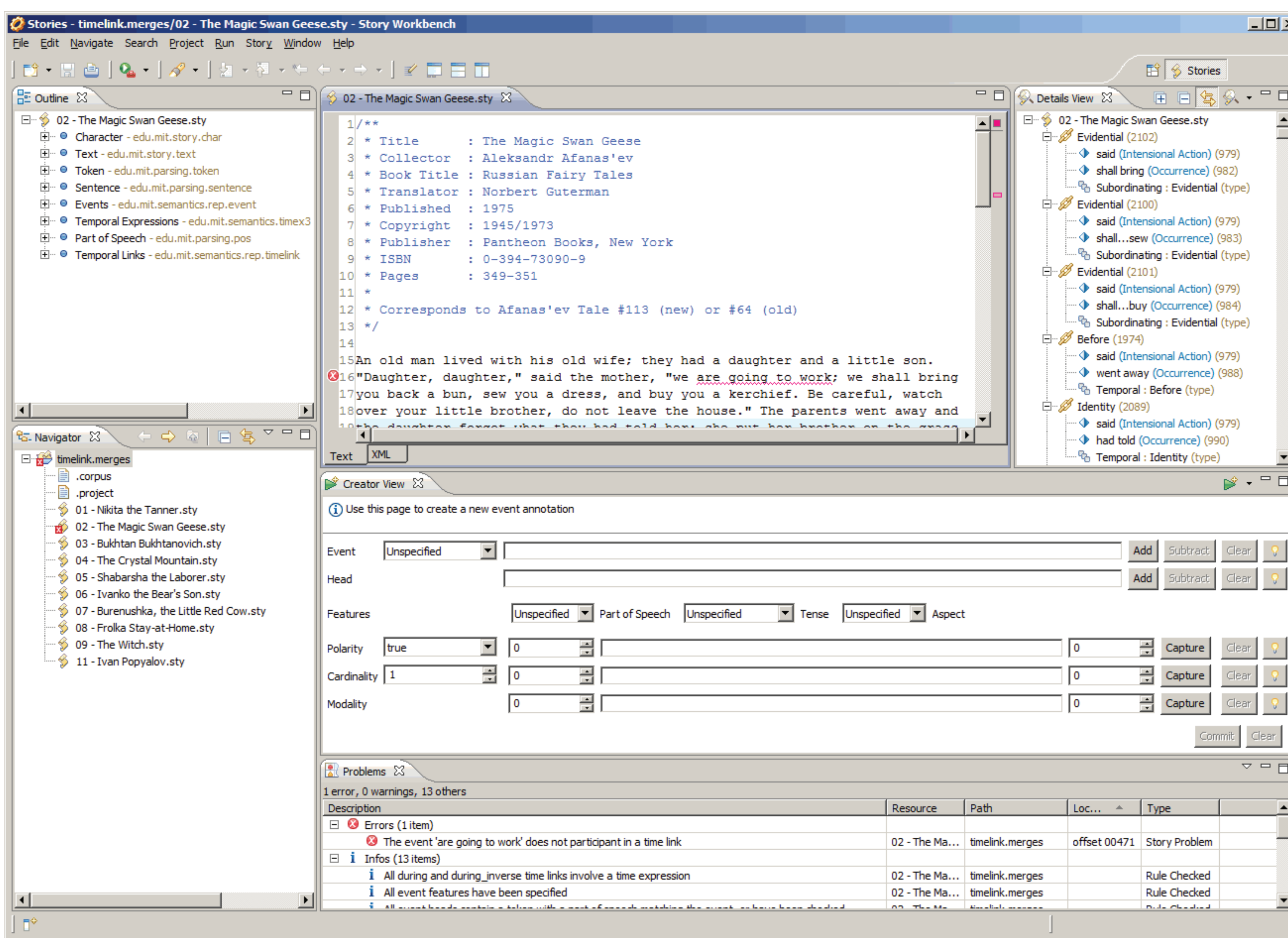
Cambridge, MA 02139 USA

## The Purpose

The purpose of this work is to demonstrate the learning, by computer, of Proppian morphological functions (Propp 1968). Later work will endeavor to test how morphologies differ across cultures, and whether people are sensitive to the presence of these functions in their cultural narratives. We begin with 16 single-move tales from Propp's original corpus (about 22k words), translated into English. The tales are then annotated semi-automatically by The Story Workbench annotation tool for 17 different layers of meaning.

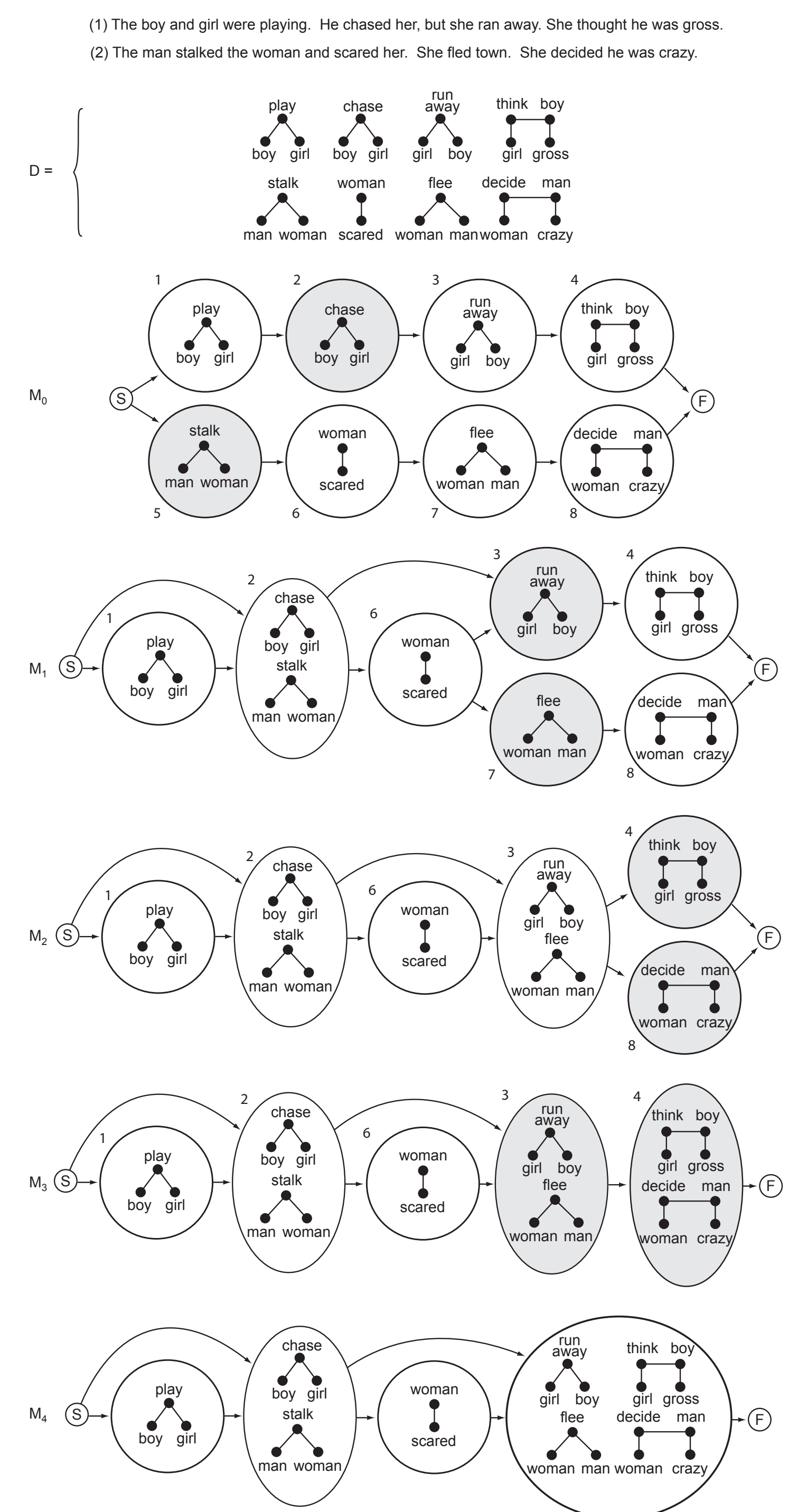
1. Tokens - location of each word token
2. Multi-word Expressions - words that are made up of multiple tokens
3. Sentences - location of each sentence
4. Part of Speech Tags - a Penn Treebank tag for each word token and multi-word expression
5. Lemmas - a lemma (i.e., stem, root form) for each word or multi-word expression not already in root form
6. Word Senses - a Wordnet sense for each token or multi-word expression
7. Context-Free Grammar Parse - a CFG parse of each sentence
8. Referring Expressions - locations of all expressions that refer to something
9. Referent Attributes - properties (unchanging attributes) of referents referred to in the text
10. Co-reference Relationships - which referring expressions refer to the same referent (co-refer)
11. Time Expressions - location, type, and value of temporal expressions, as defined by TimeML
12. Events - location, features, and type of event mentions, as defined by TimeML
13. Temporal Relationships - event-event, event-time, or time-time temporal relationships, as defined by TimeML
14. Referent Relationships - event-event, event-referent, or referent-referent unchanging non-temporal relationships
15. Semantic Roles - predicate features and arguments, as defined in PropBank
16. Mental State - mental state valencies as consequences of actions, as described by Lehnert *et al.*
17. Proppian Functions - locations of functions as identified by Propp

## The Story Workbench (Finlayson 2008)



## Analogical Story Merging (Finlayson 2009)

Analogical Story Merging (ASM) is an algorithm for extracting Proppian functions from semantically annotated text. ASM is a variation of Bayesian Model Merging. The figure below illustrates its operation. First, stories (1,2) that have been transformed into structured representations ( $D$ ) are used to construct an initial model ( $M_0$ ) that can generate all and only the stories observed. The algorithm then searches the space of state merges, where two states in the model (corresponding to two events in the story) are merged. The posterior, as calculated by Bayes' rule, directs the search to the optimal model (in the example,  $M_4$ ).



## Future Work

Of great interest will be to run ASM over sets of folktales from different cultures, and examine the differences.

Also of great interest is whether cultural participants are sensitive to the functions extracted from their own culture's folktales. I have proposed several experimental paradigms to test this.

## Evaluation Metrics

I will use at least three metrics to evaluate the output of ASM over Propp's single-move tales.

- (1) I will create a synthetic morphology and test how reliably ASM can extract it from synthetic data. This approach has already led to a demonstration of the ability of ASM to extract plot units like *Revenge* and *Pyrrhic Victory* that have been intentionally embedded in tales.
- (2) I will compare Propp's own analysis with ASM's. For this purpose, I am also annotating the locations of Propp's functions in his tales.
- (3) Cross-validation analyses using different subsets of the input data.



1st Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS) workshop  
 21 October 2010; Vienna, Austria

