Journal of Universal Computer Science, vol. 17, no. 2 (2011), 276-295
submitted: 17/9/09, accepted: 7/12/10, appeared: 28/1/11 © J.UCS

Let Me Tell You a Story – On How to Build Process Models

João Carlos de A. R. Gonçalves
(NP2Tec – Research and Practice Group in Information Technology
Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil
joao.goncalves@uniriotec.br)

Flávia Maria Santoro
(NP2Tec – Research and Practice Group in Information Technology
Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil
flavia.santoro@uniriotec.br)

Fernanda Araujo Baião
(NP2Tec – Research and Practice Group in Information Technology
Department of Applied Informatics, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil
fernanda.baiao@uniriotec.br)

Abstract: Process Modeling has been a very active research topic for the last decades. One of its main issues is the externalization of knowledge and its acquisition for further use, as this remains deeply related to the quality of the resulting process models produced by this task. This paper presents a method and a graphical supporting tool for process elicitation and modeling, combining the Group Storytelling technique with the advances of Text Mining and Natural Language Processing. The implemented tool extends its previous versions with several functionalities to facilitate group story telling by the users, as well as to improve the results of the acquired process model from the stories.

Keywords: Knowledge Management, Computer-supported collaborative work, Text mining, Business Process Modeling
Categories: I.2.7, I.7, H.4

1 Introduction

Process modeling has been adopted by organizations in various contexts, including supporting the suggestion for best practices review and for information systems engineering, as well as a basis for defining strategies for Information Technology (IT) and systems design [Weston et al., 04]. Process modeling is also fundamental for IT architecture planning. Recent studies suggest the Service-Oriented Architecture - SOA as a promising solution in this area, with processes being used as the starting point for service identification [Woodley and Gagnon, 05].

Despite its relevance in different contexts, process modeling is still a costly and complex task. The typical approach involves an analyst who conducts interviews with the performers of the tasks [Castano et al., 99; Lin et al., 02] and represents the

extracted knowledge through models. In this case, the quality of the models strongly depends on the ability of the responsible analyst and on the individuals that reported the scenarios of the process [Hickey and Davis, 04; den Hengst and de Vreede, 04].

Another approach for process elicitation that has been discussed in literature is process mining [Aalst et al., 05], which aims at extracting information from information systems event logs (such as ERP – Enterprise Resource Planning or BPMS – Business Process Management Systems) aiming to discover the process model from its implementation, trying to identify how the procedures are actually (and computationally) performed. Those techniques capture activities that are either automatic or intensively computer-supported. However, a business process may also incorporate human and interactive activities, which will never be present in system logs. Thus, strictly considering logs for mining activities may not lead to a complete business model with all its nuances.

Santoro, Borges and Pino [Santoro et al., 10] proposed the use of Group Storytelling based on the experience of collaborative processes elicitation proposals [Santoro et al., 00; Freitas et al., 03; Dennis et al., 03; den Hengst and de Vreede, 04]. Their approach suggested the collaborative narrative technique, where the information about the activities of everyday life are collected through stories told by its performers, who describe their work, the difficulties encountered, as well as suggestions for solving common problems.

This paper proposes an extension to the approach previously presented in [Gonçalves et al., 09] and [Gonçalves et al., 10] by applying the group narrative technique supported by a groupware tool, in association with text mining techniques and natural language interpretation for the automatic generation of process models from stories. The results are workflow models that help process model designers to deal with distributed knowledge besides carrying out necessary abstractions. This approach addresses the mentioned problems found both in interview-based and process mining approaches.

The paper is organized as follows: Section 2 describes related work, proposals for business processes discovery; Section 3 presents the method, which is the basis of this proposal, Section 4 details the stage of text mining to achieve model process, Section 5 presents results from two case studies, highlighting the automatic knowledge acquisition outcome, and Section 6 concludes the paper.

2 Related Work

Process discovery has been addressed in the literature through the use of both human and automatic approaches. Human approaches typically apply traditional elicitation techniques, such as interviews, to collect relevant information from all the roles involved in the process execution. Automatic approaches encompass computer-supported knowledge discovery techniques from the information system execution logs, which has been called process mining [Aalst et al., 05].

The main disadvantages of interview-based approaches are the possible biased selection of the interviewed people and the analyst's own bias when directing the interviews, users omitting mentioning "obvious" information, interviewers having poor understanding of the problem domain, the vague expression of performed activities, and conflicting perceptions from different interviewees. Some problems are

comparable to software requirements elicitation, where a number of methods [Hickey and Davis, 04] have been proposed.

On the other hand, automatic approaches try to build process models with less human involvement. The process model is based on business information and it subsequently guides the design and implementation of information systems. These systems, in turn, record important information about the activities and events being executed in the event logs.

The main constraint regarding process mining and other automatic approaches for business modeling is that they require that process activities are intensively computer-supported and, moreover, that all relevant data regarding the process execution is digitally recorded. Therefore, they are not able to capture activities which are executed without a system intervention.

Recent studies [Ingvaldsen et al., 05] [Ingvaldsen, 06] [Ghose et al., 07] [Sinha et al., 08] point to another approach for automatic process discovery. Instead of system logs, they focus on plain text process descriptions and related documents, relying on the application of Natural Language Processing and Text Mining techniques for information extraction. This enables automatic process elicitation from documents such as interview reports, which commonly occur in organizations practice.

Text mining (TM) is a process in which a user seeks to extract useful information from a document collection, by using a suite of analysis tools. Traditionally, the TM process encompasses 3 phases: pre-processing, pattern discovery, and pattern visualization.

Preprocessing techniques handle the identification and extraction of representative models for natural language documents. Several text processing techniques for text mining are originated from the area of Information Extraction (IE). IE seeks to extract semantic information from documents. There are 4 basic types of semantic information that may be extracted: entities, attributes, facts, and events. IE may be seen as a limited form of complete text comprehension [Feldman and Sanger, 07].

A story is a natural way to transmit and share knowledge. It has been successfully used in several contexts to make knowledge explicit [Leal et al., 04] [Santoro and Brézillon, 06] [Schäfer et al., 04]. Using both natural language (text) and contextual elements (categorization of parts of the story) [Leal et al., 04], story tellers can express their experience in work processes. Stories have the advantage of reproducing the situations associated with their contexts - the knowledge that is difficult to capture in interviews or through extraction from information systems. Since collectively told, a story incorporates a range of perspectives. Business processes instances can also be viewed as stories played by individuals who perform specific roles depending on the circumstances [Santoro et al., 10].

Alvarez [Alvarez, 02] conducted a study about knowledge elicitation interviews and discourse analysis, and pointed out that, during the course of knowledge elicitation interviews, a “storytelling frame” can be witnessed on the interviewees. This pattern, rich in knowledge, is therefore resisted by the analyst, who brings the process back to the usual turn-taking question-answering model.

Also, in [Schütt, 03], storytelling is considered as a relevant way to disclose and propagate critical knowledge within an organization, making it an interesting proposal for Knowledge Management applications.

The proposed approach argues that text mining and information extraction techniques may be applied on a repository of stories, reported in natural language, to extract useful information for building processes models. The method proposed in [Santoro et al., 10] is reviewed and an automatic model extraction stage is added.

3 A Method for Process Elicitation Based on Collective Stories

Santoro, Borges and Pino [Santoro et al., 10] proposed a method for business processes elicitation based on group storytelling. In this proposal, actors of a process collaboratively describe their way of acting through stories. The method follows three stages that start from concrete facts told by participants, then build abstractions and classify these facts, and end with a workflow model. This method was refined to include two automatic steps, in which algorithms for workflow mining are applied on the story texts [Gonçalves et al., 09; 10]. [Fig. 1] depicts the extended proposal.

There are three essential roles involved in the execution of the method: teller, facilitator and modeler. Tellers are the individuals who participate in the process and, consequently, can make their activities explicit through a story. The facilitator is an experienced professional who provides support in connecting the facts and producing coherent stories. The modeler is an analyst who refines the generated graphical model based on abstractions drawn from the stories.

In the first stage, groups of tellers are selected; they should tell stories, as freely as possible, related to the various situations in their day-to-day activities (process instances). This stage is supported by a collaborative tool, named ProcessTeller.

The second stage is responsible for examining told stories in order to extract process models elements: activities, flow, events, business rules, entries and exits. In this stage we apply text mining algorithms [Oliveira, 08] to identify the basic elements (activities, actors, events), as well as logical and temporal relationships among them (flows), facilitating workflow generation.

Finally, in the third stage of the method, the elements of identified process will be converted into models in a graphic notation, such as the Business Process Modeling Notation. Several alternative models can be generated (according to the findings in the previous step). These models will be presented to participants (who play the modeler role), who may combine them and, finally, validate and generate the final version of the process model.

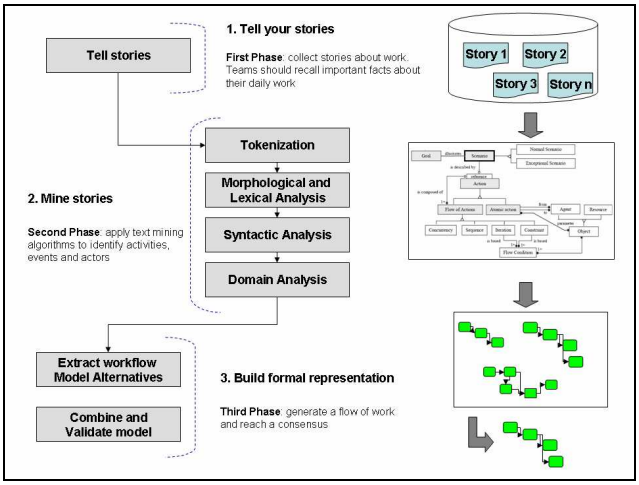

Figure 1: Process Elicitation method [Gonçalves et al., 09]

3.1 A Process-Oriented Tool for Group Storytelling

Group Storytelling has at its core the idea of free-form writing and an almost "chaotic" flow of thoughts and ideas, since it is a collaborative technique based on narratives detailed as freely as possible. The proposed method focuses on process elements extraction from the stories and supports the process analysts during the modeling task. Story Mining is more oriented to a specific goal (the process itself) than the original storytelling technique.

We developed the ProcessTeller application based on Tellstory groupware [Leal et al., 04]. It supports a narrative structure composed of events (fragments of the story) disposed on a flow, characters (who participate at the events), documents (files related to the story) and support functionalities, such as the creation of comments and voting, used for interaction between users and conflict-solving during storytelling.

In ProcessTeller, each user is identified by a login and is assigned as the moderator of the stories he/she creates. Users can also ask to participate in another user's story. In order to bridge the gap between the freedom of expression of storytelling technique and the necessity of focusing in process elicitation in order to filter specific knowledge, a few modifications and new functionalities were introduced. The goal was to enhance the efficiency of the method, while preserving the core idea and the advantages of the Group Storytelling technique. These functionalities are described in next sections.

3.1.1 Classification of Narratives Using Story Groups

Originally, TellStory allowed users to add stories at the same set of narratives. By doing so, stories from different subjects (payment descriptions, fairy tales, user reports, among others) were manipulated as if they were of the same group. This can seriously hamper any attempt of automatic processing of the stories' repository, as the

common problems of text mining (usually defined as "noise" data) would be amplified.

A directory structure of stories organized by theme was implemented, in order to mitigate this problem.

3.1.2 Document Linking to Story Events

The selection of story excerpts from the first phase of the method is made based on the "domain terms" present at each part of a story's text at the repository. This is an important component of our approach, and is directly related to its efficiency. It is possible for tellers to attach documents that are related to each event of the story.
Each event is associated to its related files. The proposed Story Mining method can then use the selected files as a source of domain knowledge specific for the event to be processed.

3.1.3 Event-Level Character Detection

The identification of process actors is an important part of the elicitation and modeling tasks, since they indicate which activities should be represented in the final model. In a similar way, the Story Mining method uses them to draw attention to the parts of a story, usually a sentence or a paragraph, that tends to provide relevant knowledge. In order not to hinder the free-form characteristic of the Group Storytelling technique, the detection of characters inside a story event description was preferred to any other intrusive proposal – such as forcing the user to describe characters for each event, for example. Therefore, when the user tries to save a story event, the tool verifies if the event description does not refer to any of the previous registered characters of the story. If so, it asks the user if he/she might want to register a new character. Again, it only notifies and suggests instead of compelling the user to do something.

3.1.4 Starting Event for each Story

Due to the goal-oriented nature of a process model, a new field, named "Starting Event" was implemented to explicitly enable the indication of the beginning of the story at the moment of its creation. The user can describe the facts that enabled the story to happen. Again, this improvement was proposed in order to make the narrative become a little more process-oriented, while taking additional care not to interfere with the storytelling process.

4 Applying Text Mining to Collaborative Stories

Automated extraction of models is a very complex task, especially when taking unstructured text as input. Writing stories in a free style presents several advantages which were already discussed, but also poses known obstacles for natural language processing, such as ambiguity and lack of clarity. Free text does not have, by definition, the requirements for workflow generation, since structural elements may be "mixed" in the same story event and manual knowledge gathering from texts could

represent as much work for process designers as using traditional techniques, such as conventional interviews.

Thus, the solution presented in our approach is a combination of the storytelling model with the representation of scenarios proposed by [Achour, 98]. This approach preserves the richness of stories and at the same time allows the extraction of useful knowledge to describe a business process.

4.1 Correspondences between the Scenario Models, Stories and Processes

The notion of a process element, which is the focus of Story Mining knowledge acquisition, can be defined as a generic activity, composed by an actor and its corresponding actions. This definition guides the second phase of the method and the CREWS model can be used for further correspondence between the process elements and parts of a story. The CREWS model [Achour, 98] was adapted here.

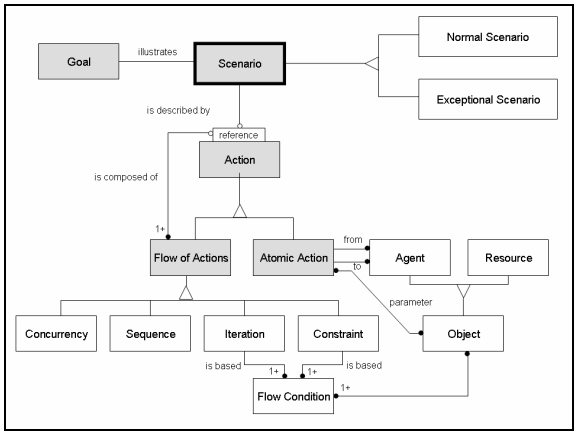

Figure 2: Scenario Model [Achour, 98] adapted for process model

[Fig. 2] shows the CREWS Model, with Scenario being the central element. It can be defined as a "behavior limited to a possible set of interactions with a purpose, occurring between different actors" [Achour, 98]. Thus, each scenario is limited to a Goal to be achieved. Also, the scenario can be “Normal”, in which everything occurs as expected; or “Exceptional”, when an unexpected event happens, altering the predicted course of action. At first, an analogy between actors of a scenario and the characters in a story seems simple. However, a user can describe various actions involving several players in the same story event, thus generating complex candidate workflows from each story event. A simple example of a complex event description is depicted in [Fig. 3].

```
Goal: Product payment
Agents: Clerk, Buyer
Actions:
1. The clerk explains to the buyer the
   types of payment accepted
2. The buyer selects one of the types of
   payment offered
3. If the payment involves credit card then
   i. The clerk asks the buyer to put the credit
      card on the teller machine.
   ii. The buyer inserts and enters his code.
4. The clerk gives the receipt to the buyer
5. The buyer receives his product.
```

Figure 3: Example Scenario for Product Payment

There is a similarity between the description of a scenario and a business process. Each is composed of actors (characters who perform actions) and actions (activities that are performed by actors). Even more, some process instances are similar to "Normal Scenarios", in which the flow of activities happens normally. Other instances, however, may be considered exceptions of the day-to-day process instances, which are related to the "Exceptional Scenario" concept as well.

Following this rationale, a collaborative story describing a business process is analogous to a scenario as well. The Goal is the intention that guides the process and its activities, and can also be considered as part of the rationale behind the story. Different stories – expressing different viewpoints about a single fact– are also similar to the "Normal" and "Exceptional" classification of the scenario model.

Moreover, general syntactic text structures may correspond to the basic elements of the scenario model. Two elements can be highlighted as the basic building blocks of a scenario (and a process as well), namely Objects and Actions.

Objects are defined as Noun Phrases (NPs), which describe either Agents (who perform Actions) or Resources (that suffer the effects of an Action). They may also refer to the story characters and the process roles, although they are broader than this classification, entailing scenario elements such as computer systems, equipments, etc.

An Action can be defined as a Verb Phrase (VP), describing activities being performed during the stories. The most common type of Action is the Atomic Action, a form of simple action between Agents and Objects, describing the activities being performed at the story by the characters. Another type of Action is the Flow of Actions, made of one or more Actions with a different flow than the common Atomic Action. At story-level, they describe constraints and conditions happening during the events being told. They will be grammatically defined as a number of Actions and Objects linked together by conjunctions, adverbs or other grammatical elements that will enable a higher degree of complexity during the description of facts.

These elements of grammar ("trigger words" that imply a Flow) can be of four different types: Concurrency (actions running in parallel to each other), Sequence (actions running on a sequential manner), Iteration (multiple executions of an action) and Constraint (a condition stated for the action to occur). They can be divided into two main groups: Concurrency and Sequence are basic modifications of the flow of activities. Iteration and Constraint depend on other Actions, as conditions and

parameters for its flow of activities. Some examples of trigger words are depicted below:

While the patient rests, the doctor prepared the vaccine (Concurrency)

After the workers leave, the maintenance process begins (Sequence)

For each equipment built, quality control must verify the output (Iteration)

If there are no more issues to solve, the system goes offline (Constraint)

Trigger words may be managed as part of a repository, such as a thesaurus or taxonomy. Its nature, however, is different, as they form the basic means for detecting flows of actions in a general linguistic way. Also, as a logical conclusion from the correspondences defined above, additional criteria for classification of Actions can be established: An Action that does not have a trigger word on its structure is an Atomic Action. With the establishment of a relationship between the scenario model, narratives and process models, the Story Mining method can be described and implemented, using the guidelines defined for its execution. The need of "capturing" the actions and agents of an event implies that the techniques of natural language processing and/or text mining should be thoroughly analyzed in order to find a solution to be implemented for this problem. To do so we adopted several steps (as shown at the second phase in [Fig. 1]). In order to illustrate each step, example story events are presented below, with the domain words highlighted in italics:

Event 1: "The *system* generates an *estimating* template consisting of the phases, activities and tasks selected for the *project* or *project* phase."

Event 2: "When planning complete *projects* the *estimating* is typically done at the activity level, using the list of tasks in the work breakdown as input to the *estimating* process."

Event 3: "However *estimators* will likely add an itemized list of *system* functions and other deliverables, to facilitate *estimating* the construction phase."

Event 4: "Multiple *estimators* prepare *estimates* for each component, compare their *estimates*, and arrive at a final *estimate* for each item."

Event 5: "The *estimates* are recorded in the *estimating* template. The template produces a summary of effort by role within the *project* organization, to assist in costing. "

Event 6: "The *system* uses the completed *estimating* template to update the metrics portion of the *project* profile to reflect the measures of *project* scope (function point counts, numbers of externals, number of trainees, etc.) used as the basis for *estimating*."

4.2 The Mining Approach

The process of Text Mining, composed by the application of techniques from Information Extraction, Information Retrieval and Natural Language Processing, can be perceived as a scientific workflow [Oliveira et al., 08]. Due to the great variety of algorithms to be used at each step of the mining phase of Story Mining, a Scientific Workflow Management System was used to support the definition and execution of the proposal, as well as allowing the repetition of the same workflow at different case

scenarios and to add flexibility on the customization of the techniques applied for the evaluation of the method itself. The workflow is depicted at [Fig. 4].

Three main parts are highlighted at the workflow. Part 1 involves the application of Information Retrieval techniques (TF/IDF) on the story's related documents (Part of the Tokenization phase), in order to provide a list of words to be used as a filter at the story events processing. Part 2 comprises the remainder of the Tokenization phase of the method, the filtering of story sentences using Part 1 output, as well as Morphological and Lexical Analysis and Syntactic Analysis. The third and last phase comprises the conversion of the textual elements into XPDL and the XML log of the method, defined as Domain Analysis at the Story Mining method. All these steps are going to be described in detail at the next sub-sections.

4.2.1 Tokenization

At this phase, the basic components of the story texts are extracted. Those are paragraphs and phrases containing relevant information of the future workflow. The selection criteria for candidate phrases are indicated by words that represent concepts related to the target domain (Words highlighted in italics at the example). This "filtering" process, when properly done, improves the process efficiency by selecting which parts of the text will be used in the following phases.

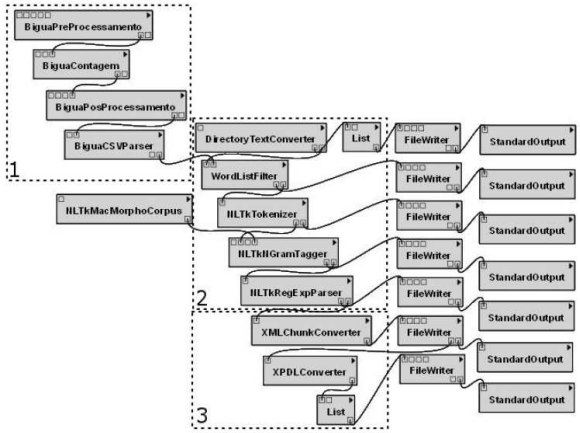

Figure 4: VisTrails Scientific Workflow for the second phase of the method

The words representing the target domain can be extracted from documents related to the stories using Text Mining techniques (as applied in other disciplines, such as Medicine [Holzinger et al., 08] and Topic Maps generation from text [Biemann et al., 03]), as well as from other structured forms of external knowledge such as ontologies, topic maps, thesauri and dictionaries.

The entire mining process can be executed without the associated documents, with a greater risk of redundant or unnecessary information, but still remaining valid for a story that contain only events for example. This initial filtering phase parses the

story's content and defines the scope of the information retrieval process. This is critical for the efficiency of the whole method in acquiring relevant process elements.

After the filtering process, the *WordPunctTokenizer* algorithm from NLTk was used for the tokenization task, dividing the story events' contents into tokens like words and additional textual elements like commas for future processing [Fig. 5].

```
['The', 'estimates', 'are', 'recorded', 'in', 'the',
  'estimating', 'template', '.', 'The', 'template',
  'produces', 'a', 'summary', 'of', 'effort', 'by',
  'role', 'within', 'the', 'project',
  'organization', ',', 'to', 'assist', 'in',
  'costing', '.', 'The', 'system', 'uses', 'the',
  'completed', 'estimating', 'template', 'to',
  'update', 'the', 'metrics', 'portion', 'of',
  'the', 'project', 'profile', 'to', 'reflect',
  'the', 'measures', 'of', 'project', 'scope', '(',
  'function', 'point', 'counts', ',', 'numbers',
  'of', 'externals', ',', 'number', 'of',
  'trainees', ',', 'etc', '.)', 'used', 'as', 'the',
  'basis', 'for', 'estimating', '.']
```

Figure 5: Example output of the PunctTokenizer

4.2.2 Morphological and Lexical Analysis

Text excerpts processed at the previous stage are now analyzed, beginning with the separation of the constituents composing each phrase. In this case, the process can include the stemming of words, removal of ambiguity present at text on the grammar level, detection of upper-case and lower-case words, syntactic tagging of words based on training corpora, shallow parsing, among other techniques. The structuring of the text, enabling the automatic extraction of information, is the main focus of this phase.

Shallow parsing, based on a training corpus selected for the application, is a very useful technique for this phase. For NLP techniques, the Trigram Tagger [Bird and Loper, 04] is applied for word-level classification of the extracted text excerpts, enabling further processing and extraction. This n-gram tagging technique checks for patterns on n-uples of words. In the case of the Trigram Tagger, it will check each triple of words from the story event.

In order to improve efficiency, supplementary algorithms are used together with the Trigram Tagger. In case the triple pattern fails to classify a word, it continues checking for patterns composed by pairs of words (Bigram Tagger), then exact match of words (Unigram Tagger) and, in the worst case, it assigns the most frequent tag of the corpus to the unknown word ("N" , meaning a Noun).

Tagging algorithms commonly use a source of external knowledge for training. One of the most common is the training corpus, usually grammatically annotated, either automatically or manually by linguistic experts. The choice for a corpus for our implementation was the Brown University Standard Corpus of Present-Day American English [Kucera and Francis, 67], being one of the most famous and extensive (more than one million words) English language corpora available.

The Brown Corpus contains newspapers, famous literary works, magazine articles, making it able to classify correctly a variety of words from both formal and

informal language. This is useful for Story Mining, as it brings the corpus closer to the form of expression of the story.

For the Portuguese language, another corpus was found, being similar to the Brown Corpus in content and aims. The MAC-MORPHO Corpus was thus chosen for applications (like the case studies presented in the paper) using Portuguese as the main language for the stories. Other corpora must be similarly found, if the Story Mining method is applied to other languages. The initial classification and tokenization of words will be the first pre-processing task for the process elements extraction [Fig. 6].

```
[[('The', 'DET'), ('estimates', 'N'), ('are', 'V'), ('recorded',
   'VN'), ('in', 'P'), ('the', 'DET'), ('estimating', 'VG'),
   ('template', 'N'), ('.', '.'), ('The', 'DET'), ('template',
   'N'), ('produces', 'VBZ'), ('a', 'DET'), ('summary', 'N'),
   ('of', 'P'), ('effort', 'N'), ('by', 'P'), ('role', 'N'),
   ('within', 'P'), ('the', 'DET'), ('project', 'N'),
   ('organization', 'N'), (',', ','), ('to', 'TO'), ('assist',
   'V'), ('in', 'P'), ('costing', 'VG'), ('.', '.'), ('The',
   'DET'), ('system', 'N'), ('uses', 'VBZ'), ('the', 'DET'),
   ('completed', 'VN'), ('estimating', 'VG'), ('template', 'N'),
   ('to', 'P'), ('update', 'V'), ('the', 'DET'), ('metrics', 'N'),
   ('portion', 'N'), ('of', 'P'), ('the', 'DET'), ('project',
   'N'), ('profile', 'N'), ('to', 'TO'), ('reflect', 'V'), ('the',
   'DET'), ('measures', 'N'), ('of', 'P'), ('project', 'N'),
   ('scope', 'N'), ('(', '('), ('function', 'N'), ('point', 'N'),
   ('counts', 'VBZ'), (',', ','), ('numbers', 'N'), ('of', 'P'),
   ('externals', 'N'), (',', ','), ('number', 'N'), ('of', 'P'),
   ('trainees', 'N'), (',', ','), ('etc', 'N'), ('.)', 'N'),
   ('used', 'VN'), ('as', 'CNJ'), ('the', 'DET'), ('basis', 'N'),
   ('for', 'P'), ('estimating', 'VG'), ('.', '.')]
```

Figure 6: Example output of the Trigram Tagger

4.2.3 Syntactic Analysis

Based on the previously classification of words, this phase starts with the analysis of the syntactic functions of the tagged text elements as well as the discovery of sentences, trigger words and other high-level structures. It also aims to structure the text and prepare it for the extraction of entities that perform actions, action descriptions as well as relationships (temporal, causal, etc) between the first two. The main focus is on the verbal and nominal constituents of each phrase, as they are able to indicate the relationship between syntactic elements, such as subjects and verbs.

A regular expression grammar is used at this phase, applied together with a Chunking [Sang and Buchholz, 00] algorithm, based on simple syntactic classifications such as noun phrases, verb phrases and sentences. The objective here is to improve the initial structuring of text, in order to point out complex elements, including sentences, noun phrases and other elements. The final product of this phase is a tagged and structured text depicting both grammatical elements and its syntactical structure [Fig. 7]. This text is the basis for the extraction of process elements.

```
(S
    (NP The/DET estimates/N)
    (VP are/V recorded/VN)
    (PP in/P (NP the/DET estimating/VG template/N)))
  ./.
(S
    (NP The/DET template/N)
    (VP produces/VBZ)
    (NP a/DET summary/N)
    (PP of/P (NP effort/N))
    (PP by/P (NP role/N))
    (PP within/P (NP the/DET project/N organization/N))
    ,/,
    (VP to/TO assist/V)
    (PP in/P (VP costing/VG)))
  ./.
```

Figure 7: Example output of the Chunking algorithm

4.2.4 Domain Analysis

At this final stage, the multi-level structure created from the original free-form text must be "translated" to workflow format. To facilitate this task, templates are developed to attend for all types of elements composing a workflow.

The template defines the domain focus. In its current status, our proposed approach focuses on process activities, since they represent a common element of most business process models available. Also, the correspondences between the scenario model and grammar elements will be used for this task, extracting grammar patterns for Actions and Objects. In order to simplify the procedure and bring it closer to a business process notation, a template will be defined [Fig. 8]:

<ACTIVITY>: Sentence containing a Verb Phrase
<ACTOR>: Noun Phrase occurring at an ACTIVITY before an ACTION
<ACTION>: A Verb Phrase at an ACTIVITY
<PARAMETER>: Grammar elements occurring at an ACTIVITY and after an ACTION

Figure 8: Example template used for Domain Analysis

This simple template can then be used to extract two outputs of the mining process. The first is a log file, containing all the extracted activities [Fig. 9]. The second output consists of an XPDL file that can be imported and edited at a Business Process Modeling tool [Fig. 10].

```
<ACTIVITY>
        <ACTOR>The system </ACTOR>
        <ACTION>generates </ACTION>
        <PARAMETER>an estimating template consisting of the phases   activities
        </PARAMETER>
</ACTIVITY>
```

Figure 9: Example of the intermediate output structure of the method

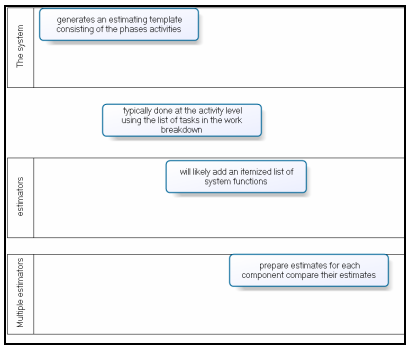

Figure 10: Example of model visualization from the XPDL file

5 Evaluation of the Proposal: Case Studies

This section describes the evaluation of the proposal as a viable method for business process elicitation and support for business process modeling. Two cases studies were performed, at different environments, including different natures of both processes to be elicited as well as different profiles of the tellers. They describe the application of the last phase of the method, as well as the results and evidence found during its application at real world scenarios.

The application of the scientific workflow was performed using different regular expression grammars for chunking, differing in their complexity level. This allowed the evaluation of the amount of relevant process information could be extracted and how it would affect the modeling task at the third phase of the method, when the modeler will have to create the final process model based on the method's output.

The first grammar (Grammar I) focused on extracting process elements, dividing and highlighting its pieces. Those pieces were separated by "trigger words" that could indicate Flows of Actions, such as Conjunctions. The second grammar (Grammar II) was simplified and focused on extracting full process elements from text.

The evaluation was performed based on the output of the workflow, using two criteria: the recognition of the extracted element as a process activity and if the element was ready for use and complete with contextual information. All the analysis was performed by the modeler, using its previously acquired experience on business process modeling.

5.1 Course Enrollment at DIA/UNIRIO

The first case study was performed at the Department of Applied Informatics of Federal University of the State of Rio de Janeiro (UNIRIO). The chosen process was "Course Enrollment" of both Master's Degree and Bachelor's Degree courses.

The tellers were selected among people involved with the process itself, including students, university professors and administrative staff of the institution. Although the amount of time necessary for a collaborative story to mature and provide useful

knowledge is a difficult thing to define, a period of time of one month was established for the storytelling phase.

During this time, eighteen tellers expressed their viewpoints using the ProcessTeller tool, adding events, characters and other story elements as well as reviewing and commenting other users' contributions.

At the end of the first phase, twenty-six story events were available for the application of mining algorithms. Seven additional related documents were also available, including course subscription form templates, process regulations, examples of calendars and lists of courses to be enrolled.

The mining workflow was then applied, using two different types of regular expressions grammars. [Tab. 1] and [Tab. 2] summarize the results, showing the number [Tab. 1] and the corresponding percentage [Tab. 2] of extracted elements, separating valid elements (those that were considered by the modeler as real process elements) from noise elements (those that were not considered as useful process elements by the modeler).

	Valid	"Noise"	Total
Grammar I	17	34	51
Grammar II	29	7	36

Table 1: Event Analysis for Case Study I

	Valid	"Noise"	Total
Grammar I	33,33%	66,67%	100,00%
Grammar II	80,56%	19,44%	100,00%

Table 2: Percentage Analysis for Case Study I

The tables depict the evaluation results, with a clear focus on the usefulness for the Business Process Elicitation and Modeling areas, as the modeler's analysis can reveal. The first grammar captured a great amount of "Noise" elements (more than 50%), while the simpler grammar (Grammar II) was able to present clearer pieces of knowledge about the process. A detailed examination of the results shows that the total of events for each grammar was different, with more elements being extracted when the search for trigger words was a criterion for extracting process knowledge.

5.2 Manage Process Elicitation at a Multinational Company

The second case study was performed at a Multinational Company, focusing on a knowledge-intensive, complex process called "Manage Process Elicitation". The available tellers were the members of the Business Process team of the organization.

After a short tutorial and hands-on training of one day, with the objective of making all tellers familiar with the Process Teller tool and the aims of the Story Mining method, the first phase of the method started. Storytelling was performed in parallel with their day-to-day activities.

The group of tellers was defined as eight members of the team, and the case study was performed during a month, resulting in fifteen story events at the end of the

storytelling phase. An important fact was that no related document was added to the story. Therefore, the original text mining workflow was adapted to skip the filtering of text based on words extracted from them.

The application of the modified mining workflow was performed, applying the two different types of regular expression grammars (Grammar I focusing on trigger words and Grammar II focusing on whole process elements), and extracting process elements for evaluation. After the modeler's analysis, a comparison was performed and [Tab. 3] and [Tab. 4] illustrate the results.

	Valid	"Noise"	Total
Grammar I	9	11	20
Grammar II	11	3	14

Table 3: Event Analysis for Case Study II

	Valid	"Noise"	Total
Grammar I	45,00%	55,00%	100,00%
Grammar II	78,57%	21,43%	100,00%

Table 4: Percentage Analysis for Case Study II

Similarly to the first case study, Grammar II seems to be able to capture process elements in a more useful way for the process analyst (and for the usage at the third step of the Story Mining approach), even with less story events present at this narrative. Also, the noise levels and valid element frequency were similar, even being a story in different contextual conditions, distinct organizational cultures (multinational enterprise x university) and applying different algorithms (number of documents and of story events). Another detail was the closer total of process elements extracted using each grammar, pointing out a possible usage of few "Trigger Words" at the narrative.

5.3 Discussion

In our approach, the quality of the process model in terms of completeness and correctness is directly dependent on the textual material provided by the tellers of the stories. It means that, if they are motivated to write about the process and to interact with other participants, to discuss issues that might arise, doubts, or even obscure details, probably the mining techniques applied later will be better successful than if they have to handle a poor text with no facets. The application that supports the method can contribute by stimulating the collaborative behavior among participants. Some improvement, however, may be possible. For example, awareness mechanisms, such as information about who has written about the same facts, and where in the story some topic is been described could motivate people to read each other parts and complement their own. Besides, coordination features might help the facilitator to know which tellers are less participative and encourage them to write down about their experiences. Also, communication tools like chat and forums would push people to argue more.

On the other hand, if too much disordered information is provided, the storytelling method might not assist the elicitation of process. Therefore, another issue to be explored is the guidance in writing the stories properly for the goal. Enhancement is also necessary in this case. The software could interpret the text as it is been written by tellers and hints could be made available: the identification of synonyms, lack of verbs, too long or too short sentences. It could advise the teller to review his text. Also, contextual elements might also be highlighted within the events of the story. Mining techniques should be refined to and tested in a number of stories.

In the case studies conducted, we evaluated the quality of the stories by comparing the resulting process models with previous ones elaborated using the traditional interview-based approach. We noticed that the content of the stories described more details about the process than in the other case. However, since the mining techniques still produce a "noisy" model, human analysis was performed over the texts to reach this conclusion.

6 Conclusions

This paper presents the refinement of a collaborative method for designing a process model based on stories. The proposal is motivated by the assumption that telling a story is natural for people, and therefore it is a good way to capture knowledge about activities performed by people in their day-to-day activities. Building abstractions from the facts being told is not only a difficult task, but also time consuming. Thus, our proposal includes phases aiming at extracting and generating models using text mining and natural language processing techniques. The usage of a collaborative tool enhances the efficiency of the method, by improving the quality of the narratives and guiding the automatic interpretation of text while, at the same time, avoiding interference on the way users tell the story.

A general-scope, language-specific corpus such as the Brown Corpus seems sufficient for identifying the common grammatical words and the "trigger words" necessary to identify flows and more complex process elements. This fact also points to the use of a tagged corpus of the specific language on which the stories will be written. The quality of the documents present on the chosen corpus will directly affect the efficiency of the method as a whole.

The text mining techniques and external knowledge sources, with their domain-specific terms, complement the corpus' role, as a source of new terms outside the scope of general language. Even if these terms do not come with their grammatical classifications, the rules generated by the Trigram Tagger, based on the grammatical patterns present at the documents of the training corpus, have potential of lessening this problem.

Our tests detected undesirable situations in the discovered process models, such as activities without actors (as seen on the second activity at [Fig. 10] and actors that have similar meanings (Ex: "Estimators" and "Multiple Estimators"). This indicates the necessity of word-level disambiguation techniques, such as stemming, which will be investigated as future work.

The syntactic and morphological analysis phases can be refined to mitigate the "noise" level and achieve more results ready to be applied by the modeler. Tests with other types of taggers, chunking grammars and new NLP techniques will also be

made in this direction. Finally, due to the presence of pronominal anaphora in text, the usage of an anaphora resolver algorithm will be studied to improve even further the extraction process.

The XPDL format was selected for the final output file of the method due to its wide acceptance and practical use for process model exchanging between different process modeling tools. The variety of ways that workflow elements and domain-specific concepts can be expressed on a story brings up the future need of an external knowledge source, like a glossary, dictionary or ontology.

Since each teller represents his particular point of view within the stories, there is always a possibility of multiple workflows for the same business process. In this case, it will be offered a functionality to support the choice for final version workflow, with the preservation of alternative flows, depicting a process of greater complexity.

The evaluation of the model output by the modeler, a professional with experience and the criteria used (usefulness for the modeling task) shows that the method can be useful for Business Process Elicitation and as a support for Business Process Modeling. The usage of two grammars showed that a complex extraction of process elements would demand additional algorithms and possibly other sources of external knowledge, due to the additional difficulty of interpretation by the modeler. However, simplified extraction of activities achieved the objectives of the mining proposal: to support the analyst during the creation of business process models, as well as enabling easier and optimized access to the knowledge present at each story.

The trigger words extracted by the second grammar point out that further experimentation and testing using more complex forms of process element extraction can be promising, aiming for the successful extraction of detailed process elements, like the Flow of Events from the CREWS Model as well as taking the Story Mining method to a higher level of knowledge extraction.

Thus, the Story Mining method can be considered as an option for other techniques or to be used together with traditional elicitation methods like structured interviews and workshops.

Acknowledgements

The authors wish to thank CAPES and Petrobras for the sponsorship of this research. Flávia Santoro is partially supported by CNPq (Brazil) grant n. 305404/2008-3.

References

[Aalst et al., 05] Aalst, W. M. P., Weijters, A.: "Process-Aware Information Systems: Bridging People e Software through Process Technology"; Wiley & Sons (2005)

[Achour, 98] Achour, C. B.: "Guiding Scenario Authoring"; Proc. 8th European-Japanese Conference on Information Modelling and Knowledge Bases, Finland (1998), 152-171

[Alvarez, 02] Alvarez R.: "Discourse Analysis of Requirements and Knowledge Elicitation Interviews"; Proc. 35th Hawaii International Conference on System Sciences (2002), 255

[Biemann et al., 03] Biemann C., Quasthoff U., Böhm K., Wolff C.: "Automatic Discovery and Aggregation of Compound Names for the Use in Knowledge Representations"; J.UCS (Journal of Universal Computer Science), 9, 6 (2003), 530-541

[Bird and Loper, 04] Bird S., Loper E.: "NLTk: The Natural Language Toolkit"; Proc. the ACL demonstration session (2004), 214-217

[Castano et al., 99] Castano, S., De Antonellis, V., Melchiori, M.: "A methodology and tool environment for process analysis and reengineering"; Data & Knowledge Engineering, 31, 3 (1999), 253-278

[Dennis et al., 03] Dennis, A. R., Carte, T. A. and Kelly, G. G.: "Breaking the rules: success and failure in groupware-supported business process reengineering"; Decision Support Systems, 36, 1 (2003), 31-47

[Feldman and Sanger, 07] Feldman, R., Sanger, J.: "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data"; Cambridge University Press (2007)

[Freitas et al., 03] Freitas, R. M., Borges, M. R. S., Santoro, F. M., Pino, J. A.: "Groupware Support for Cooperative Process Elicitation."; IX International Workshop on Groupware (CRIWG), Lecture Notes in Computer Science, 2806 (2003), 232-246

[Ghose et al. 07] Ghose, A., Koliadis, G., Chueng A.: "Process Discovery from Model and Text Artefacts"; Proc. IEEE Congress on Services (2007), 167-174

[Gonçalves et al, 09] Gonçalves, J. C. A. R., Santoro, F. M., Baião, F. A.: "Business Process Mining from Group Stories"; Proc. 13th International Conference on Computer-Supported Cooperative Work in Design, Santiago, Chile (2009), 161-166

[Gonçalves et al, 10] Gonçalves, J. C. A. R., Santoro, F. M., Baião, F. A.: "A case study on designing business processes based on collaborative and mining approaches"; Proc. 14th International Conference on Computer-Supported Cooperative Work in Design, Shanghai, China (2010), 611-616

[den Hengst and de Vreede, 04] den Hengst, M., de Vreede, G. J.: "Collaborative Business Engineering: A Decade of Lessons from the field"; Journal of Management Information Systems, 20, 4 (2004), 85-114

[Hickey and Davis, 04] Hickey, A., Davis, A.: "A unified model of Requirements Elicitation"; Journal of Management Information Systems, 20, 4 (2004), 65-84

[Holzinger et al., 08] Holzinger, A., Geierhofer, R., Mödritscher, F, Tatzl, R.: "Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses"; J.UCS (Journal of Universal Computer Science), 14, 22 (2008), 3781-3795

[Ingvaldsen et al., 05] Ingvaldsen, J. E., Gulla, J. A., Su, X., Rønneberg, H..: "A Text Mining Approach to Integrating Business Process Models and Governing Documents"; On the Move to Meaningful Internet Systems 2005: OTM Workshops, Springer (2005),

[Ingvaldsen, 06] Ingvaldsen, J. E., Gulla, J. A.: "Model-Based Business Process Mining"; IS Management, 23, 1 (2006), 19-31

[Kucera and Francis, 67] Kucera, H. and Francis, W. N.: "Computational Analysis of Present-Day American English"; Brown University Press, Providence, RI (1967)

[Leal et al., 04] Leal, R. P., Borges, M. R. B., Santoro, F. M.: "Applying Group Storytelling in Knowledge Management"; IX International Workshop on Groupware (CRIWG), Lecture Notes in Computer Science, 3198 (2004), 34-41

[Lin et al., 02] Lin, F. R., Yang, M. C. and Pai, Y. H.: "A Generic Structure for Business Process Model"; Business Process Management Journal, 8, 1 (2002), 19-41

[Oliveira, 08] Oliveira, D.: "MiningFlow: Adding Semantics to Text Mining Workflows"; Master Dissertation (in Portuguese), COPPE/UFRJ, Brazil (2008)

[Sang and Buchholz, 00] Sang, E. F. T. K., Buchholz, S.: "Introduction to the CoNLL-2000 Shared Task: Chunking"; Proc. CoNLL-2000 and LLL-2000, Lisbon, Portugal (2000), 127-132

[Santoro et al., 00] Santoro, F. M., Borges, M. R. S., Pino, J. A.: "CEPE: Cooperative Editor for Processes Elicitation"; Proc. 33rd Hawaii International Conference on Systems Sciences, IEEE Computer Society Press (2000), 1003

[Santoro and Brézillon, 06] Santoro, F. M; Brézillon, P.: "The Role of Shared Context in Group Storytelling"; Computing and Informatics, 25, 6 (2006), 1001-1026

[Santoro et al., 10] Santoro, F. M., Borges, M. R. S., Pino, J. A.: "Acquiring knowledge on business processes from stakeholders' stories"; Advanced Engineering Informatics (AEI), 24, 2 (2010), 138-148

[Schäfer et al., 04] Schäfer, L., Valle, C., Prinz, W.: "Group Storytelling for Team Awareness and Entertainment"; Proc. 3rd Nordic Conference on Human-computer Interaction, Tampere, Finland (2004), 441-444

[Schütt, 03] Schütt, P.: "The post-Nonaka Knowledge Management"; J.UCS (Journal of Universal Computer Science), 9, 6 (2003), 451-462

[Sinha et al., 08] Sinha A., Paradkar A., Kumanan P., Boguraev, B.: "An Analysis Engine for Dependable Elicitation of Natural Language Use Case Description and its Application to Industrial Use Cases"; IBM Research Report RC24712 (2008)

[Woodley and Gagnon, 05] Woodley, T., Gagnon, S.: "BPM and SOA: Synergies and Challenges"; Proc. 6th International Conference on Web Information Systems Engineering, Lecture Notes in Computer Science, 3806, New York, NY, USA (2005), 679-688

[Weston et al., 04] Weston, R. H., Chatha, K. A., Ajaefobi, J. O.: "Process thinking in support of system specification and selection"; Advanced Engineering Informatics, 18, 4 (2004), 217-229