A Semi-automatic Approach to Identify Business Process Elements in Natural Language Texts

Renato César Borges Ferreira[1], Lucinéia Heloisa Thom[1] and Marcelo Fantinato[2]
[1]*Department of Informatics, Federal University of Rio Grande do Sul, UFRGS, Porto Alegre, Brazil*
[2]*School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil*

Keywords: Process Models, Natural Language Processing, Process Element, Business Process Management, Business Process Model and Notation, Process Modeling.

Abstract: In organizations, business process modeling is very important to report, understand and automate processes. However, the documentation existent in organizations about such processes is mostly unstructured and difficult to be understood by analysts. The extracting of process models from textual descriptions may contribute to minimize the effort required in process modeling. In this context, this paper proposes a semi-automatic approach to identify process elements in natural language texts, which may include process descriptions. Therefore, based on the study of natural language processing, we defined a set of mapping rules to identify process elements in texts. In addition, we developed a prototype which is able to semi-automatically identify process elements in texts. Our evaluation shows promising results. The analyses of 56 texts revealed 91.92% accuracy and a case study showed that 93.33% of the participants agree with the mapping rules.

1 INTRODUCTION

Public or private organizations seeking for better interaction with their customers and business partners need to offer high-quality products or services. Furthermore, they seek to achieve superior standardization and efficiency in the performance of their business processes (referred in this text as process). The automation of processes performed in an organization provides greater control over costs, time, errors and redundancy in the execution of processes (Thom, 2012; Thom et al., 2009).

A business process is a "collection of events, activities, and decision point actions, involving a number of actors and objects, which collectively lead to results that bring value to the customer" (Dumas et al., 2013). Business Process Management (BPM) is defined as a set of methods, techniques and tools to discovery, analyze, redesign, implement and monitor business processes (Weske, 2007; Dumas et al., 2013). According to (Leopold, 2013), through BPM, organizations can flexibly adapt in a continuously changing business environment. Therefore, BPM involve many improvements to the organization, such as the standardization of processes, improvement, quality and quick execution of the activities (Thom, 2012; Thom et al., 2009).

The BPM lifecycle includes six steps: *(i)* processes identification, *(ii)* discovery, *(iii)* analyze, *(iv)* redesign, *(v)* implementation and *(vi)* processes monitoring and controlling (Dumas et al., 2013) . All life cycle steps are particularly important for process automation. The process modeling can be considered one of the most important and complex steps. In this step, business processes executed in the organization are designed with the use of a graphical notation such as the Business Process Model and Notation (BPMN). (Dumas et al., 2013) describe that the process modeling is a prerequisite for analysis, redesign and automation of business processes. An incorrect process modeling compromises the next steps of the BPM project, since a correct process automation originates from a precise process modeling.

We have learned from practice that the design of a particular process (e.g., healthcare processes) can be very complex, not only due to its variety and the need for flexibility but also because they require knowledge of several domain terms (Thom et al., 2010). Moreover, it lead to ambiguities and interpretation problems between process analysts and domain experts. Process modeling comprises several methods for processes discovery, such as user's interviews, workshops, brainstorming and documents of the organization (Dumas et al., 2012). These documents

Ferreira, R., Thom, L. and Fantinato, M.
A Semi-automatic Approach to Identify Business Process Elements in Natural Language Texts.
DOI: 10.5220/0006305902500261
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 3*, pages 250-261
ISBN: 978-989-758-249-3

can have many sources, such as reports, forms, letters, notes from call centers, surveys, research, business policies, textbooks, systems knowledge management, e-mail messages, event data of information systems, web pages, texts documents and interviews.

However, such methods may have limitations due to miscommunication between analysts and users, the lack of documentation and standardization of processes, and the lack of user information. Content management professionals consider that 85% of the information in companies is stored in an unstructured way, especially as text documents (Blumberg and Atre., 2003). To develop meaningful process models, the process analyst obtains abstract information on how these processes are implemented. To create the initial process model (as-is model), the process analyst usually collects several pieces of information about the process through the mentioned methods to establish the initial understanding of the process (Dumas et al., 2013). The acquisition of the initial process model in a BPM project requires 60% of the total time spent (Herbst, 1999).

Several works (Friedrich et al., 2011; Chueng et al., 2007; Goncalves et al., 2011) have demonstrated that the extraction of process models can minimize the effort of the process analyst to capture, mainly through user's interviews process models. These interviews rarely lead to the understanding of the entire process, since they often only describe knowledge from isolated parts of the process.

In this context, this paper proposes a semi-automatic approach to identify process elements in Natural Language texts. The identification of process elements in texts assists in the construction of a processes template and can thus extract the process models from it. We observed that most of the work described in the literature considers that the texts in natural language are described in a way that only analysts are able to extract process models from that. This means that the texts are developed with very specific keywords and sentences which denote process elements. Therefore, novice process analysts cannot extract process models from it. In addition, organizations have several unstructured textual information, which can be used as possible sources of information for process design. Thus, natural language texts are mostly not prepared to be directly used by process model extraction tools. Therefore, this shows how complex is to identify process elements in texts. Thus, the approach presented in this paper not only contributes to the identification of process elements but also to inform completeness of natural language texts as well as missing process elements (e.g., start, end events, tasks, swimlanes, parallel gateways (AND) and exclusive gateways (XOR)).

We developed a prototype to semi-automatically identify process model elements in natural language texts. The tool uses as input a collection of documents such as reports, manuals, forms and norms within organizations. We combined a large set of tools from Natural Language Processing (NLP) based on the mapping rules, which were particularly developed for our approach. The evaluation of our prototype shows very encouraging results. Considering a set of 56 texts, the accuracy was 91.92% based on machine learning evaluation metrics and measures for information retrieval. Furthermore, the validation through a survey showed that most the of participants, i.e. 93.33%, agree with the mapping rules.

For the identification of process model elements in natural language texts, we used the following methodology:

- Definition of mapping rules: in this step, the mapping rules were defined and afterwards applied in natural language texts.
- Development of a prototype: implementation of a prototype to semi-automatically identify process elements in natural language texts based on the mapping rules.
- Evaluation of identified process elements: our approach follows two evaluation perspectives: First, mapping rules validation through a survey with potential users, in particular process experts. Second, prototype validation based on the set of NLP tools.

The remainder of this paper is structured as follows. Section 2 provides related works. Section 3 shows the proposed approach to identify process elements in natural language texts. Section 4 shows the evaluation and results analysis. Finally, Section 5 concludes the paper.

2 RELATED WORKS

In this section we review the most relevant works regarding process model extraction from natural language texts. The state of art can be divided in two related categories: the extraction of process models from natural language texts and text generation from process models. Table 1 provides an overview of the identified state of art.

The analyses of works exploring process model extraction from natural language texts shows two main aspects.

First, we should consider the source of information of the natural language texts. (Friedrich et al.,

Table 1: State of art for Identify Process Elements in Natural Language Texts.

Categories
Process Model Extraction from Text
- Generate Process Models From Text.
- *(Friedrich et al., 2011)*
- *(Chueng et al., 2007)*
- Process Mining from Natural Language Text
- *(Santoro et al., 2009)*
- *(Jiexun et al., 2010)*
- *(Goncalves et al., 2011)*
Text Generation from Process Models
- Generate Text from Process Models
- *(Leopold et al., 2014)*
- *(Meitz et al., 2013)*
- *(Leopold, 2013)*
- Inconsistencies Between Process Models and Text
- *(van der Aa et al., 2015)*
- *(van der Aa et al., 2016)*
- Text Structuring
- *(Heinonen, 1998)*
- *(Hearst, 1994)*
- *(Hearst, 1997)*
- *(Hynes and Bexley, 2003)*

2011) proposed by the extraction of process models from textual descriptions. The proposed approach considers three outlooks: first, syntactic analysis, determination of a syntax tree and grammatical relationships between the parts of the sentences; second, semantic analysis, extraction of the meaning of words or phrases; and third, anaphora resolution, identifying concepts that are referenced using pronouns *(we, he* and *it)* and articles *(this, that)*. In this work, the source of information from text is a limitation. For instance, the texts analyzed for the generation of process models need to be grammatically correct in the English language, i.e., it is necessary to remove and manually correct words or sentences that are grammatically incorrect, so that the text becomes grammatically correct. Furthermore, the text must not contain questions and needs to be described sequentially. In order to solve this problem, in a previous work (Ferreira and Thom, 2016), we conducted an introductory approach to generate process-oriented text from natural language, from this approach we concluded that natural language text must be processed before the extraction of process models. In particular, (Chueng et al., 2007) describe that the source of information from text are heterogeneous information sources (e.g., corporate documentation, web-content, code etc.).

The second aspect refers to process models extracted from natural language texts. (Santoro et al., 2009) and (Goncalves et al., 2011) described an approach that explores the narrative technique associated with text mining and natural language interpretation for generating process models. The paper shows that miscommunications can occur, e.g. each author represents their individual point of view within the stories, there is always a possibility of multiple workflows for the same business process (Wfmc, 2005). Therefore, the source of information can have ambiguities. (Jiexun et al., 2010) proposed a process mining framework named policy-based process mining (PBPM) for the automatic discovery of process models based on business policies. Considering that policy texts is a new topic in BPM research and text mining, the approach requires additional research efforts to be entire validated and produce practical solutions. Thus, there is a small training set and a small portion of positive examples in the approach.

Regarding text generation from process models, we identified three main approaches from the works we analyzed. (Leopold, 2013) described an approach for generating natural language texts from process models. In this approach, the author describes challenges to generate texts from process models such as text planning; sentence planning; surface realization and flexibility. The limitations of this work refer to the fact that the sentences generated are comparatively short and elementary. Another limitation is to ensure a stable level of complexity of the texts created manually so it would be necessary to train the text classifiers. Finally, the modeling process is not well documented since the text generated is not structured. To reduce the time and effort needed between process model and textual description, (van der Aa et al., 2015) and (van der Aa et al., 2016) describe an approach to identify inconsistencies between a process model and a corresponding textual description. This approach can be used to identify process models in a collection that are likely to diverge from their accompanying textual descriptions.

To reduce *inconsistency* or *ambiguity* of process models extracted from natural language texts, a *text structuring* is necessary. Many researches seek to identify how to optimally structure natural language texts using paragraphs. Similarity metrics such as the semantic relatedness between words to compute the lexical cohesion between the sentences of a text are implemented by many methods (Hearst, 1994; Hearst, 1997; Morris and Hirst, 1991). Therefore,

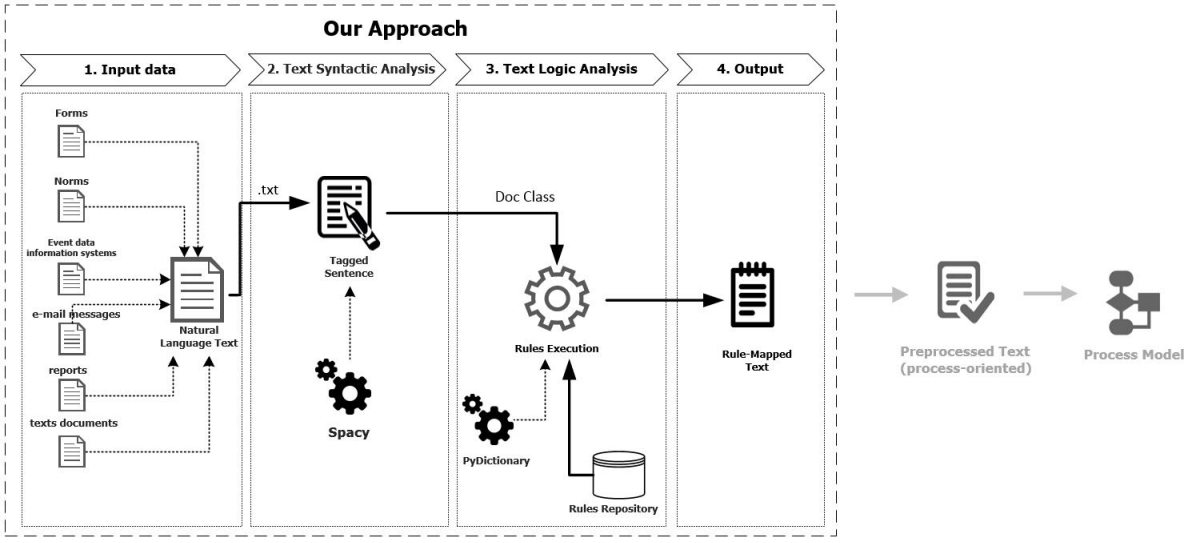


Figure 1: A semi-automatic approach to identify process elements in natural language texts.

a text can be heuristically subdivided into multiple paragraphs. More approaches seek to use the similarity distribution for identifying the optimal fragment boundaries (Heinonen, 1998). (Hynes and Bexley, 2003) shows that paragraphs containing more than 100 words are less understandable than paragraphs with fewer words.

3 AN APPROACH TO IDENTIFY PROCESS ELEMENTS IN NATURAL LANGUAGE TEXTS

In this section, we present an approach to identify process elements in natural language texts. In our previous research (Ferreira and Thom, 2016), we conducted an introductory approach to generate process-oriented text from natural language. That research serves as the foundation for the approach presented in this paper.

Figure 1 shows our approach to identify process elements in natural language texts. This approach consists of four main steps including input data, text syntactic analysis, text logic analysis and output. In the following sections, we introduce and explain each of these steps.

3.1 Input Data

In the English language, there are many classifications of texts. Each classification has different characteristics, such as words, phrases, and issues related to each text particularity. Examples of classifications include texts descriptor; comparison and contrast; order of importance; problem and solution; cause and effect; sequential. In sequential texts, the information is organized in steps or process and is explained in the order they occur[1].

The characteristics presented in the sequential text has similarities with business process models. There are keywords common in sequential texts, such as *first*, *second*, *near*, *then*, *finally*, *following*, *now*, *after*, among others. These words show possible relationships (correlations) with modeling elements of BPMN such as activities, swimm, gateways, pools, swimlanes, etc.

In this context, text documents can have many sources, such as: forms, norms, event data of information systems, e-mail messages, etc. These sources are also called as natural language text, including sentences not structured that do not correspond to the sequential texts and hence make very difficult the extraction of process models from it. The output of this step are data with *.txt* format separated by sentences.

3.2 Text Syntactic Analysis

In order to obtain a tagged sentence is necessary to consider the syntactic analysis parsing from text documents. The purpose of the syntactic analysis is to determine the structure of the input text. Per (Allen, 1995), analyze the syntactical parser, we need to perform three aspects: first, a parser has as input a sentence and as a result produces the analysis; second, a grammar has a set of rules that the parser can use; and

[1] http://www.ereadingworksheets.com/text-structure/; last accessed 2016-11-11

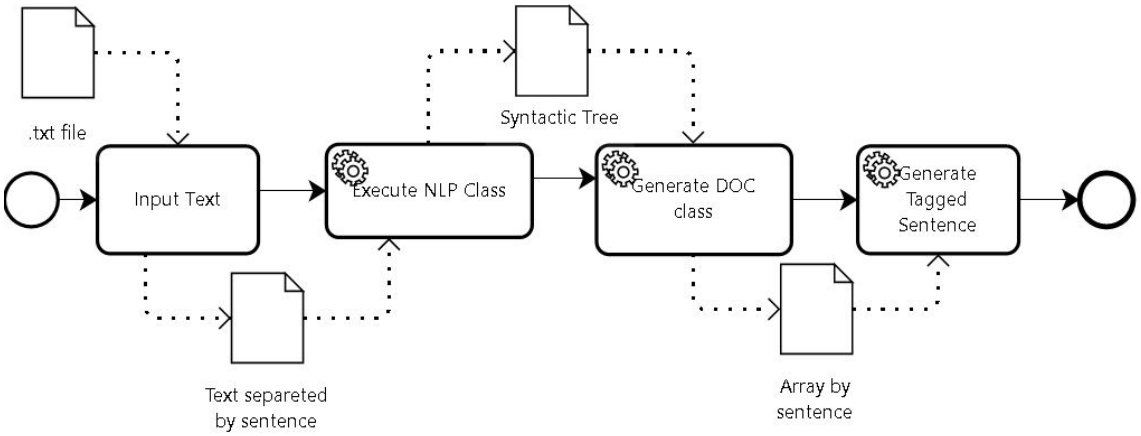


Figure 2: Structural overview of the step of text syntactic analysis in order to generate a tagged sentence.

third, a lexicon, which is a dictionary of legal words and their parts of speech (e.g., *verb, adverb, adjective, subject, direct object, indirect object* etc.). Part of speech tags provides significant information about the role of a word in its narrow context. It may also provide information about the inflection of a word (de Kok and Brouwer, 2011). There are many tools refering to parts of speech (POS-taggers). Examples of POS-taggers are the Brill tagger (Brill, 1992), GATE [2], RASP system (Briscoe et al., 2006), and NLTK[3].

For our approach, we selected as syntatic parser Spacy[4]. The selection of *Spacy* was based on accuracy (Choi et al., 2015), and its supporting for all the requirements of our prototype development (e.g., parts of speech).

In this work, syntactic parser and parts of speech are an important factor to identify process elements in sentences. In order to identify parts of speech of sentences in *Spacy*, we need to divide all sentences from text, the result is a *txt* file separated by sentence as shown in task *"input text"* of Figure 2. Afterwards, to achieve all the syntactic analyze of the sentence, it is necessary to execute a *NLP* class which is identified and related in a syntactic tree (as shown in the *"execute NLP class"* service task in Figure 2). Such tree contains all words related to morphological classes. Subsequently, the parser generates a *DOC class* shown in a service task *"generate DOC class"* of Figure 2. Such class is an array of the object with the number of positions equivalent to the number of words in the sentence, where each position is a word of the sentence that would be handled in the next step (text logic analysis). Furthermore, each position contains all features (e.g., tokenization, sentence recognition, part of speech tagging, lemmatization, dependency parsing, and named entity recognition) of the word on the text. Finally, tagged sentence is generated with all the sentences analyzed by the syntactic parser. Figure 2 shows a structural overview of the steps.

3.3 Text Logic Analysis

In order to support and minimize the effort of the process analyst on the modeling step, we developed a set of mapping rules and word correlations to identify process elements in natural language texts.

The rules originate from a diverse set of grammatical classes (part of speech), such as *verb, pronoun, article, numeral, and noun*. Based on the study of the grammatical classes, there is no pattern describing the way the grammar classes should be presented in the text. This shows that they are related to each other and represent some process elements. For instance, the sentences that contains *subject*, *verb* and *object* represent a process element, such as: *manual task* and there is no dependence on the order in which they occur in the text. In some sentences, grammar classes merge among themselves, for example, sentences containing the order of the *verb*, followed by a *subject* and subsequently by an *object* representing a *manual task* of the process and they are positioned in different ways in the sentence.

Mapping rules have been defined manually, and each rule is classified according to a category of the basic set of BPMN modeling elements (e.g., flow objects, connecting objects, swimlanes etc.) proposed by OMG (Object Management Group) which are recurrent in business processes. The complete set includes 33 mapping rules. From these, nine refers to activities (manual tasks and service tasks), ten to

[2]https://gate.ac.uk/; last accessed 2016-11-12
[3]http://www.nltk.org/; last accessed 2016-11-12
[4]https://spacy.io/; last accessed 2016-11-12

Table 2: Rules for identification of primary activities.

Activities – primary rules		
Rules	**Description**	**Sentence example**
Rule 1	<***subject***>+ <***verb***>+ <***object***>	*The Support Officer* <**subject**> *updates* <**verb**> all *group calendars* <**object**>
Rule 2	<***subject***>+ <***aux***>+ <***verb***>+ <***object***><**in the future)**	*The secretary* <**subject**>*will* <**aux**>*send*<**verb**>to *dispatch* <**object**.>
Rule 3	<***verb***>+<***article***>+ <***object***>	- *choose* <**verb**>*a* <**article**>*document* <**object**>. - it *do* <**verb**>*a*<**article**>*order* <**object**>.
Rule 4	<***subject***>+<***verb***>+ <***object***>+ <***conjunction***>+ <***verb***>+ <***object***>	*A client* <**subject**>*calls* <**verb**>the *help desk* <**object**>*and* <**conjunction** >*makes* <**verb**>a *request* <**object**>.
Rule 5	<***object***>+<***subject***>+ <***verb***>	The *severity* <**object**>of the *claimant* <**subject**>is *evaluated* <**verb**>.
Rule 6	<***subject (occult)***>+ <***verb***>+ <**conjunction** >+<***verb***>+ <***object***>	The first activity is to *check* <**verb**>*and* <**conjunction**>*repair* <**verb**>the *hardware* <**object**>.

Table 3: Rules for identification of primary events.

Events – primary rules		
Rules	**Description**	**Sentence example**
Rule 1	<***subject***>+ <***verb***>+ <***object***>	*After the agent* <**subject**>has *confirmed* <**verb**>the *claim* <**object**>to the clerk.
Rule 2	<***subject***>+<***verb***>+ <***agent*** >+ <***object***>	The *SCT physical* <**subject**>file was *stored* <**verb in the past**>*by* <**agent**>the *Back office* <**object**>. (passive voice)
Rule 3	<***object***>+ <***verb present perfect***>	*...Urgent document* <**object**>has been *received* <**verb**>by *the Manager. ...*
Rule 4	<***object***>+<***verb past***> + <***subject***>	...a *message* <**object**>was *generated* <**verb**>to the *customer*<**subject**>.

Table 4: Rules for Identification of Primary Exclusive Gateway (XOR).

Exclusive gateways (XOR) – primary rules		
Rules	**Description**	**Sentence example**
Rule 1	<***verb***>+ <***signal word***>+ <***subject***>+ <***object***>	It first *checked* <**verb**>*whether* <**signal word**>*the claimant* <**subject**>is *insured* <**object**>by the organization.
Rule 2	<***signal word***>+ <***condition***>+ <***task/event***>+<***alternative signal word***>+ <***task/event***>	*If* <**signal word**>*the claimant requires two or more forms* <**condition**>, *the Department of customer select the forms* <**task**>. *Otherwise* <**alternative signal word**>, *Department of customer it requires documentation* <**task**>.
Rule 3	<***task/event***>+ <***signal word***>+ <***condition***>	*After that they enter into a firm commitment to buy the stock and then offer it to the public* <**task**>, *when* <**signal word**>*they haven't still found any reason not to do it* <**condition**>.
Rule 4	<***task***>+ <***signal word***>+ <***condition***>+<***alternative signal word***>+<***task***>	*The clerk checks* <***task***>*whether*<***signal word***>*the beneficiary's policy was valid at the time of the accident* <**condition**>. *If not* <**alternative signal word**>, *it send to Department of the intelligence* <**task**>.

events, four to parallel gateways (AND), seven to exclusive gateways (XOR) and three to swimlanes. Therefore, the mapping rules proposed in this paper relate to categories of flow objects and swimlanes. The labels of the tasks and events are based on the sentence, there will be at least one *subject*, *verb* and *object*. According to (Mendling et al., 2010) and (Mendling, 2013), the activity labels are represented by the *verb* and *object* as for example *"Inform Complainant"*. Therefore, the use of labels is an important process modeling guidelines because it directly affects the clarity and understanding of the process.

The mapping rules were defined in two categories: primary rules constituted by frequency in natural language text and they represent a category of the basic BPMN modeling elements; and secondary rules which were identified with less frequency in texts. In this paper, we introduced only primary rules. The secondary rules is available at https://goo.gl/kpdEeF.

In the context of this research, for identifying process elements in sentences, we observed the sentences containing verbal tenses in the present or future, which represent activities. On the other hand, sentences that contain verbal tenses in the past or present perfect of the English Language represent events. Another difference is the elaboration of event labels. According to (Mendling, 2013) and (Leopold, 2013), event labels are represented by an *object* of the

Table 5: Rules for identification of primary parallel gateways (AND).

Parallel Gateway (AND) – primary rules		
Rules	**Description**	**Sentence example**
Rule 1	<***task/event***>+ <***signal word***>+ <***task/ event***>	*Forward the document* <**task**>, *In parallel with this* <**signal word**>, *the RCC shall also notify the Executive Board* <**task**>.
Rule 2	<***signal word***>+ <***task***>+ <***conjunction***>+ <***task***>+ <***task***>	*In parallel with this* <**signal word**>*Department of sell send the document* <**task**>*and* <**conjunction**>*notify the department of engineering* <**task**>. Then, *the document is processed* <**task**>.
Rule 3	<***signal word***>+ <***task/event***>	*In the meantime* <**signal word**>, *the engineering department prepares everything for the assembling of the ordered bicycle* <**task**>.

Table 6: Rules for identification of swimlanes.

Swimlanes		
Rules	**Description**	**Sentence example**
Rule 1	**The subject of the sentence.**	<***subject***>perform <***task/event*** >
Rule 2	<***task*** >+ <***indirect object*** >	*She then submits an order* <**task**> to the *customer* <**indirect object**>.
Rule 3	<***event*** >+ <***indirect object*** >	*The Manager forwarded the form* <**event**> to *Official* <**indirect object**>.

sentence and followed by a *verb* in the present participle like *"Invoice Created"*.

For instance, rule five from Table 2 provides an example to identify activities in sentences. The rule contemplates the sequence of an *object*, followed by a *subject* and afterwards a *verb*. The process modeling of this sentence would become one task with the label is the merge of the *verb* and *object*. Therefore, such sentence will be a candidate for process modeling. The rest of the rules for this process element (activities) follows the same pattern of identification (see Table 2).

Rule two from Table 3 illustrates an example to identify events in sentences. The rule considers the sequence of a *subject*, followed by a *verb*, afterwards by an *agent*[5] and finally by an *object*. Thus, such sentence will be a candidate for process modeling. The rest of the rules for this process element (events) follows the same pattern of identification (see Table 3).

The mapping rules from Table 4 and Table 5, describes words that denote control flow. Such words are denominated as *signal words* and *alternative signal words*. Such words refer to a condition[6] in the sentence (conditional clause). These words are divided into two groups:

- Parallel gateways (AND): words that refer to parallelism
  - signal words: while, meanwhile, in parallel, concurrently, meantime, in the meantime, in parallel with this, in addition to, simultaneously, at the same time, whereas.
- Exclusive gateways (XOR): words that refer to exclusion
  - signal words: if, whether, if not, or, in case [of], otherwise, either, only, till, until (unless), when, only if.
  - alternative signal words: but, then, else, or, unless, without, either, otherwise, other, if its is not, otherwise.

In order to identify synonyms of these *signal words* and *alternative signal words*, we implemented a Python module to get synonyms of words. Such module referred to *PyDictionary*[7].

For instance, rule two from Table 4 provides an example to identify exclusive gateway (XOR) in sentences. The rule considers the sequence of a *signal word*, followed by a *condition*, afterward a *task* or *event*, followed by an *alternative signal word* and finally by a *task* or *event*. For this reason, such sentence will be a candidate for process modeling. The rest of the rules for this process element (XOR) follows the same pattern of identification (see Table 4).

The rule two from Table 5 describes an example to identify parallel gateway (AND) in sentences. The rule consider the sequence of a *signal word*, followed by a *task*, then by a *conjunction*[8], afterwards by a *task* and finally a *task* once more. Therefore, such sentence will be a candidate for process modeling. The rest of the rules for this process element (AND) follows the same pattern of identification (see Table 5).

[5]An agent is the complement of a passive verb that is the surface subject of its active form. In our approach, the preposition "by" is included as a part of agent.

[6]Conditions are defined such as task or events in our approach

[7]https://pypi.python.org/pypi/PyDictionary; last accessed 2016-11-14

[8]A conjunct is a dependent of the leftmost conjunct in coordination. The leftmost conjunct becomes the head of a coordinated phrase.

Finally, the mapping rules from Table 6 further illustrate an example to identify swimlanes from the sentence. We observed that *subject* or *indirect object* will always be the swimlanes of the sentence. Furthermore, the *subject* can be a human being, equipment, system or something that practices the action.

In summary, we applied the mapping rules in the sentences to identify process elements. Thereafter, rule-mapped text is generated with all the sentences analyzed by mapping rules.

3.4 Output

In order to obtain a rule-mapped text, it is necessary to analyze all sentences of the text. Such sentences can be understood as a candidate to extract process models from it according to the mapping rules identified.

In summary, to describe process elements in natural language texts this step can identify: *start events*, *end events*, *swimlanes*, *actions*, *tasks*, *task labels*, *events labels* and show the number of process elements in the analyzed text. In addition, to inform completeness of natural language texts as well as missing process elements.

Figure 1 shows the next step in this approach, described as preprocessed text (process-oriented). Such text is defined as a structure that allows to identify: the participant associated with an activity; swimlanes associated with each pool; interaction between pools (*message flow*); events (*start*, *intermediate* and *end*) and control flows (*parallel gateways*, *inclusive gateways* and *exclusive gateways*). It is expected to generate a template of how the text should be structured for the extraction of process models from the text. In other words, our approach is a prerequisite for generating preprocessed text (process-oriented).

4 EVALUATION

In this paper, we conducted two experiments to demonstrate the feasibility of our approach. The first one refers to the validation of the mapping rules. To do so we performed a survey with potential users, in particular process experts. Second, the prototype validation was based on a set of NLP tools.

We used a survey strategy to interact with potential users and obtain information considering their experiences including mainly those ones in the process experts (e.g., chief process officer, business engineer, process designer, process participant, process owner etc) and from the BPM area. The only requirement to answer the survey was basic knowledge in experience in BPMN. Thus, the survey was applied using *Google Forms*[9]. The form was available from Oct 20 to Nov 10, 2016. We advertised it in social networks and websites, consequently, 43 answers were collected from participants, including process experts, software developer, students, among others. The survey was divided into three steps.

The first step aimed to gather general information on the participants' background, including: profession, education, experience in BPM, amount of experience time in BPM, experience in BPMN, amount of experience time in BPMN, knowledge in process modeling guidelines and knowledge in the grammar of the English Language.

In the second step, the goal was to get opinions on the participants about the identification of which process elements could be identified in the sentence considered in the survey. The purpose of this step is to enable the agreement of the answers according to the mapping rules created for process element shown in the sentence. The sentences are:

1. *A customer brings a defective computer and the CRS checks the defect and hands out a repair cost calculation back.*
2. *If the customer decides that the costs are acceptable, the process continues, otherwise, she takes her computer home unrepaired.*
3. *The ongoing repair consists of two activities, which are executed, in a parallel order. The first activity is to check and repair the hardware, whereas the second activity checks the software and configures the hardware.*

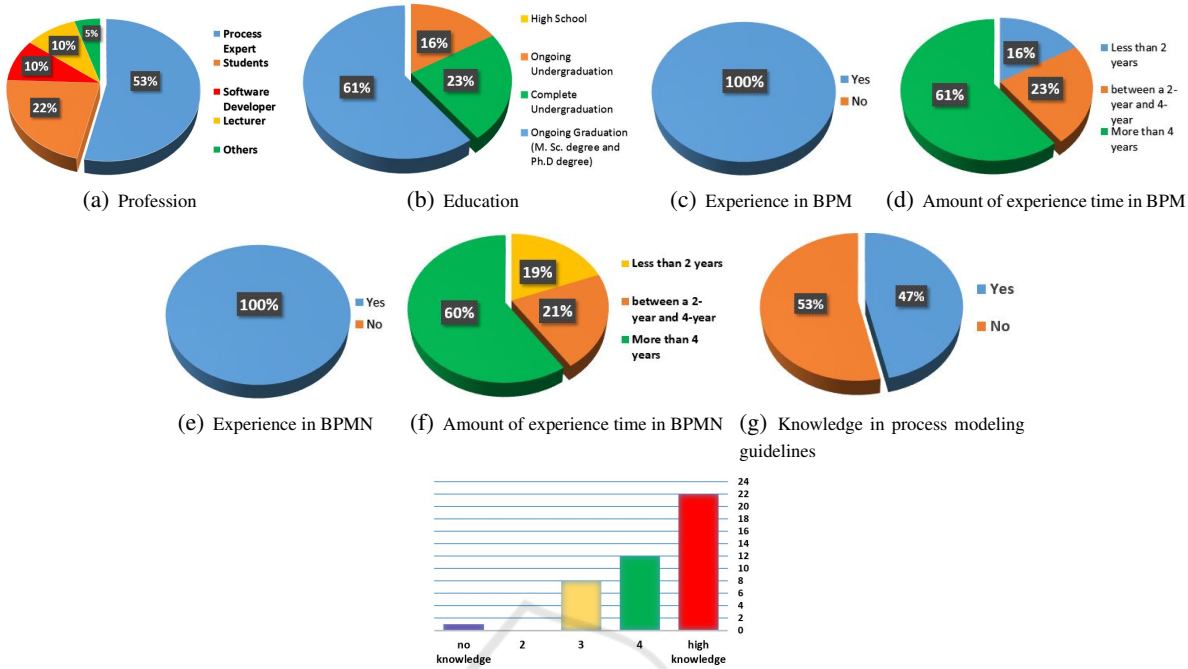
In the third step, our goal was to get an opinion of the participants about the process modeling shown in the survey. Thus, six process models were created from sentences shown in the survey, and only two process models were modeled according to the mapping rules. The rest of the process models were purposefully modeled incorrectly. The reason is to verify whether the participants' answers are in accordance with the mapping rules for such sentence.

In terms of results obtained with the three steps of the survey, Figure 3, 4 and 5 shows all the related data for each step respectively.

The evaluation conducted in this paper demonstrated encouraging results. In order to get all the information given by process experts, we analyzed the answers of 22 participants. We have selected the follows characteristics of participants:

1. Process experts
2. Experience in BPMN

[9]https://www.google.com/forms/about; last accessed: 2016-11-17



(a) Profession (b) Education (c) Experience in BPM (d) Amount of experience time in BPM

(e) Experience in BPMN (f) Amount of experience time in BPMN (g) Knowledge in process modeling guidelines

(h) knowledge in grammar of the English Language

Figure 3: Results obtained from the first step of the survey with all participants. Figure3(a) shows that more than half of the participants are process experts (53%), 22% are students, 10% are software developer and lecturer and finally 5% are other professions. Figure 3(b) illustrates that 61% of the participants hold ongoing graduation (M.Sc. and Ph.D. degree), 23% hold complete under graduation, 16% hold ongoing under graduation. Figure 3(c) describes that 100% of the participants have experience in BPM. Figure 3(d) shows that 61% of the participants have more than four years of experience in BPM, 23% have between two and four years and 16% have less than two years. Figure 3(e) shows that 100% of the participants have experience in BPMN. Figure 3(f) illustrates that 60% of the participants have more than four years of experience in BPMN, 21% have between between two and four years and 19% have less than two years. Figure 3(g) shows that 53% of the participants have knowledge in process modeling guidelines and 47% no knowledge. Finally, Figure 3(h) shows that 22 participants have a high knowledge in English Language, 12 have great knowledge, eight have good knowledge and only one have no knowledge.

3. More than two years of experience in BPMN

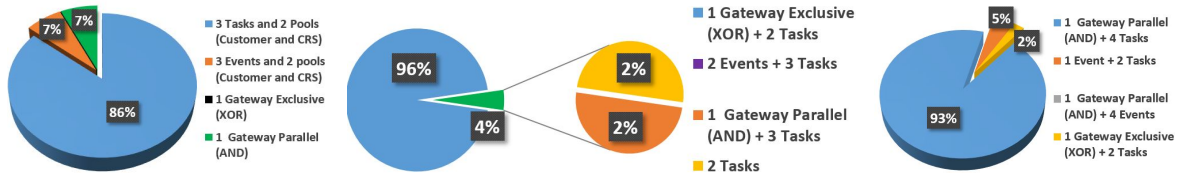
The second step of the survey shows that 90% of the participants agree with the model presented in the first sentence of the survey. Regarding the second sentence 100% of the participants agree with the modeling while 90% agree with the modeling in the third sentence. All sentences were modeled based on the proposed modeling rules by our approach.

The third step of the survey also demonstrates encouraging results. The results obtained from process modeling of the Figures 5(a), 5(c), 5(e) and 5(f), which represents purposefully incorrectly modeling according to the mapping rules proposed by our approach, the majority of the participants selected disagree (90%, 100%, 95% and 77% respectively) with the proposed modeling. On the other hand, the second and fourth modeling (Figures 5(b) and 5(d)), which represent correct modeling according to the mapping rules, the majority of the participants agree (68% and 81% respectively) with the proposed modeling.

In this study, we conducted one experiment to demonstrate the feasibility of our prototype. Our experimental study includes four sets of natural language text: the first set is from the BMW owner's Manuals & Documents[10], the second set is from the Immigrant Visa Process[11], the third set is from Federal Network Agency of Germany[12], and the fourth set is from Vista Project Office Documentation Plan[13]. In total, the set of natural language text contains 387 sentences in 56 texts. Altogether, we found that 140 sentences represents activities, 106 events, 98 exclusive gateways (XOR) and 43 parallel gateways (AND) (see Table 7).

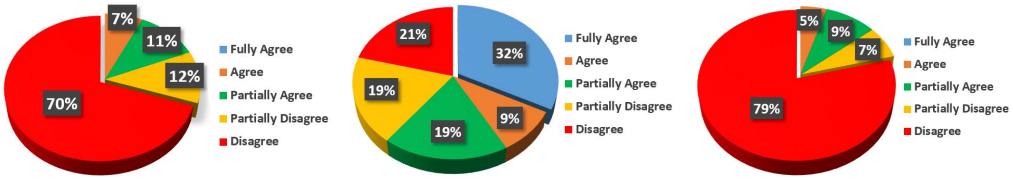
We used standard machine learning evaluation

[10]https://goo.gl/REUmu6; last accessed 2016-11-12
[11]https://goo.gl/rPLqXE; last accessed 2016-11-12
[12]https://goo.gl/KuQOBw; last accessed 2016-11-12
[13]https://goo.gl/MxzAAH; last accessed 2016-11-12

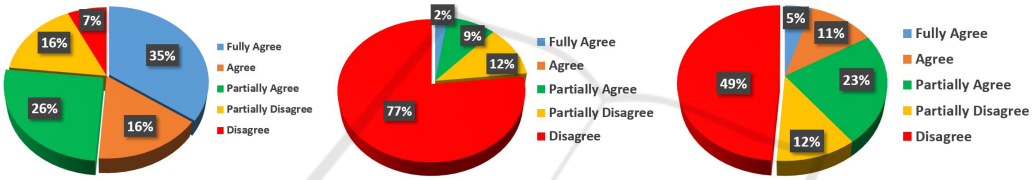


(a) First sentence (b) Second sentence (c) Third sentence

Figure 4: Results obtained from the second step of the survey with all participants. The results of the first sentence shown in Figure 4(a) describes that 86% of the participants agreed with the modeling. Figure 4(b) illustrate that 96% of the participants agreed with the modeling. Finally, Figure 4(c) describes that 93% of the participants agreed with the modeling. In this step of the survey, all sentences were modeled according to the mapping rules proposed by our approach.



(a) First modeling (b) Second modeling (c) Third modeling



(d) Fourth modeling (e) Fifth modeling (f) Sixth modeling

Figure 5: Results obtained from the third step of the survey with all participants. For the Figures 5(a), 5(c), 5(e) and 5(f) which represents purposefully incorrectly modeling according to the mapping rules proposed by our approach, the majority of the participants disagree (70%, 79%, 77% and 49% respectively) of the proposed modeling. On the other hand, the process modeling Figures 5(b) and 5(d) are represented according to the mapping rules proposed by our approach, the majority of the participants agree (60% and 77% respectively) of the proposed modeling.

metrics and measures for information retrieval (Japkowicz and Shah, 2011; Forbes, 1995; Manning et al., 2008), precision (equation 1), recall (equation 2), accuracy (equation 3) and F-measure (equation 4), to evaluate the performance of our prototype. Accuracy measures the overall correctness. Precision, recall, and F-measure evaluate the correctness for each class (activities, events, parallel gateways and exclusive gateways). F-measure is the harmonic mean of precision and recall (Jiexun et al., 2010). Such metrics can be calculated as follow:

- Number of correctly identified instances (η);
- Total of number of instances (χ);
- Number of correctly identified instances for class i (β);
- Total number of instances identified as class i (τ);
- Total number of instances in class i (Φ).

$$Precision(i) = \frac{(\beta)}{(\tau)} \quad (1)$$

$$Recall(i) = \frac{(\beta)}{(\Phi)} \quad (2)$$

$$Accuracy = \frac{(\eta)}{(\chi)} \quad (3)$$

$$F-measure(i) = \frac{2 \times precision(i) \times recall(i)}{precision(i) + recall(i)} \quad (4)$$

Table 7 summarizes the performance of our prototype. We report precision, recall, accuracy and F-measure values for the four process elements: activities, events, parallel gateways and exclusive gateways.

Class activity achieved 83.57% precision, 78% recall, 80.68% F-measure and 85.52% accuracy. Class event reached 93.39% precision, 81.14% recall, 86.84% F-measure and 92.24% accuracy. Class exclusive gateways (XOR) describes that achieved 72.44% precision, 93.42% recall, 81.60% F-measure and 91.73% accuracy. Class parallel gateways (AND)

Table 7: Results of performance of our prototype proposed by our approach.

Class	Found	Precision	Recall	F-Measure	Accuracy
Activities	140	83.57%	78%	80.68%	85.52%
Events	106	93.39%	81.14%	86.84%	92.24%
Exclusive Gateway (XOR)	98	72.44%	93.42%	81.60%	91.73%
Parallel Gateway (AND)	43	88.37%	97.43%	92.68%	98.44%
Total	387	84.44%	87.49%	85.45%	**91.92%**

presents that obtained 88.37% precision, 97.43% recall, 92.68% F-measure and 98.44% accuracy. In general, our prototype achieves higher performance.

4.1 Evaluation Analysis

This section brings the results of a survey which was developed with the aim of demonstrate the users opinion regarding our prototype.

The results of the survey show a great acceptance of the mapping rules by process experts (22 participants). For instance, in the second stage of the survey, the acceptance was on average 93.33%. Therefore, it shows the feasibility of applying the mapping rules to identify process elements in texts.

For prototype allowed us to evaluate the performance. Despite the small number of sentences, we can see through evaluation metrics very promising results in terms of mainly accuracy, precision and recall. On average 91.92%, 84.44%, 87.49% respectively. Therefore, the prototype would allow semi-automatically identify process elements in natural language texts.

5 CONCLUSIONS

In this paper, we proposed a semi-automatic approach to identify process elements in natural language texts. We have created 33 mapping rules to identify process elements in the texts. In addition, we have developed a prototype to semi-automatically identify process elements in texts. We combine a large set of tools from NLP based on the mapping rules. The evaluation of our prototype which was based on a set of 56 texts presented 91.92% of accuracy. Furthermore, the validation through the survey demonstrated that 93.33% of the participants agree with the mapping rules. Thus, our approach minimizes the effort of the process analyst to capture business process elements from natural language texts and indicates completeness of the texts based on BPMN rules.

Despite these promising results, one limitation of our approaches refers to the automatically generation of rules. Hence, we intend to generate them through artificial intelligence. Although our approach contains a category of the basic BPMN modeling elements, in future works we will explore the creation of mapping rules for other BPMN process elements, such as message flows, sub-processes, exception flows, data object, sequence flow and inclusive gateways (OR), etc. In addition, Figure 1 shows the next step in this approach, described as preprocessed text (process-oriented). It is expected to generate a template of how the text should be structured for the extraction of process models from text. Our approach can be considered as a prerequisite for generating preprocessed text (process-oriented).

REFERENCES

Allen, J. (1995). *Natural Language Understanding*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.

Blumberg, R. and Atre., S. (2003). The problem with unstructured data. DM Review.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, COLING-ACL '06, pages 77–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Choi, J. D., Tetreault, J. R., and Stent, A. (2015). It depends: Dependency parser comparison using A web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 387–396.

Chueng, A., Koliadis, G., and Ghose, A. (2007). Process discovery from model and text artefacts. *2007 IEEE Congress on Services*, 00(undefined):167–174.

de Kok, D. and Brouwer, H. (2011). *Natural Language Processing for the Working Programmer*.

Dumas, M., Rosa, M. L., Mendling, J., Mäesalu, R., Reijers, H. A., and Semenenko, N. (2012). Understand-

ing business process models: The costs and benefits of structuredness.

Dumas, M., Rosa, M. L., Mendling, J., and Reijers, H. A. (2013). *Fundamentals of Business Process Management.* Springer.

Ferreira, R. C. B. and Thom, L. H. (2016). An approach to generate process-oriented text from natural language. page 77. XII Brazilian Symposium on Information Systems.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring*, 11(3):189–206.

Friedrich, F., Mendling, J., and Puhlmann, F. (2011). Process model generation from natural language text. pages 482–496.

Goncalves, J. C. A., Santoro, F. M., and Baião, F. A. (2011). Let me tell you a story - on how to build process models. volume 17, pages 276–295. Journal of Universal Computer Science.

Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

Heinonen, O. (1998). Optimal multi-paragraph text segmentation by dynamic programming. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 1484–1486, Stroudsburg, PA, USA. Association for Computational Linguistics.

Herbst, J. (1999). An inductive approach to the acquisition and adaptation of workflow models. In *Proceedings of the IJCAI'99 Workshop on Intelligent Workflow and Process Management: The New Frontier for AI in Business*, pages 52–57.

Hynes, G. and Bexley, J. (2003). Understandability of banks' annual reports. In *69th Association for Business Communication Annual Convention*, pages 1–11, Albuquerque.

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, New York, NY, USA.

Jiexun, L., Wang, Jiannan, H., Zhu, and Leon, J. (2010). A policy-based process mining framework: mining business policy texts for discovering process models. *Information Systems and E-Business Management*.

Leopold, H. (2013). *Natural language in business process models*. Springer.

Leopold, H., Mendling, J., and Polyvyanyy, A. (2014). Supporting process model validation through natural language generation. volume 40. IEEE Transactions on Software Engineering.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Meitz, M., Leopold, H., and Mendling, J. (2013). An approach to support process model validation based on text generation. volume 33, pages 7–20.

Mendling, J. (2013). *Managing Structural and Textual Quality of Business Process Models*, pages 100–111. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mendling, J., Reijers, H. A., and van der Aalst, W. M. P. (2010). Seven process modeling guidelines (7pmg). *Inf. Softw. Technol.*, 52(2):127–136.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48.

Santoro, F. M., Goncalves, J. C. A., and Baiao, F. A. (2009). Business process mining from group stories. *International Conference on Computer Supported Cooperative Work in Design*, pages 161–166.

Thom, L., Reichert, M., and Iochpe, C. (2009). Activity patterns in process-aware information systems: Basic concepts and empirical evidence. *International Journal of Business Process Integration and Management (IJBPIM)*.

Thom, L. H. (2012). *Gerenciamento de Processos de Negócio e Aplicabilidade na Saúde e na Robótica*. Biblioteca Digital Brasileira de Computação.

Thom, L. H., REICHERT, M., IOCHPE, and C. OLIVEIRA, J. P. (2010). Why rigid process management technology hampers computerized support of healthcare processes. WIM - X Workshop de Informática Médica.

van der Aa, H., Leopold, H., and Reijers, H. A. (2015). Detecting inconsistencies between process models and textual descriptions. In *International Conference on Business Process Management*, pages 90–105. Springer.

van der Aa, H., Leopold, H., and Reijers, H. A. (2016). *Dealing with Behavioral Ambiguity in Textual Process Descriptions*, pages 271–288. Springer International Publishing, Cham.

Weske, M. (2007). *Business Process Management: Concepts, Languages, Architectures*. Springer-Verlag, Berlin.

Wfmc, W. M. (2005). Wfmc: Process definition language: Xpdl 2.0. page 164.