Revising the Vocabulary of Business Process Element Labels

Agnes Koschmider[1](✉), Meike Ullrich[1], Antje Heine[2], and Andreas Oberweis[1]

[1] Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany
{agnes.koschmider,meike.ullrich,andreas.oberweis}@kit.edu
[2] Institut für Deutsche Philologie, Ernst-Moritz-Arndt-Universität, Greifswald, Germany
antje.heine@uni-greifswald.de

Abstract. A variety of methods devoted to the behavior analysis of business process models has been suggested, which diminish the task of inspecting the correctness of the model by the process modeler. Although a correct behavior has been attested, the process model might still not be feasible because the modeler or intended user is hampered in her comprehension (and thus hesitates e.g., to reuse the process model). This paper addresses the improvement of comprehension of process element labels by revising their vocabulary. Process element labels are critical for an appropriate association between the symbol instance and the real world. If users do not (fully) understand the process element labels, an improper notion of the real process might arise. To improve the comprehension of element labels algorithms are presented, which base on common hints how to effectively recognize written words. Results from an empirical study indicate a preference for such revised process element labels.

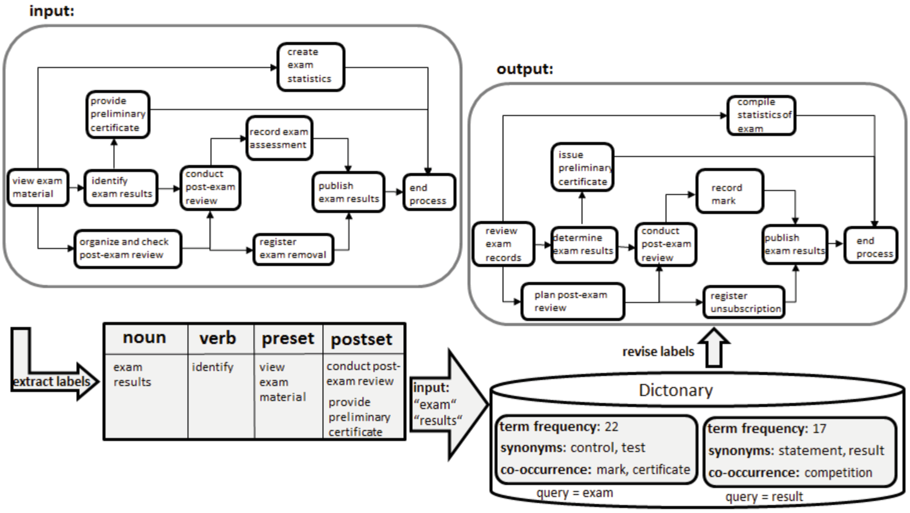
1 Introduction

The labeling of business process model elements is still a mainly manual task and requires a great deal of experience of the process modeler. Highly skilled process modelers tend to find easier (and better) labels for process model elements than modeling beginners, who also might omit activities or might have problems to find an appropriate abstraction level for activities [1]. Process element labels are critical for an appropriate association between the symbol instance and the real world [2]. If users do not (fully) understand the process element labels, an improper notion of the real process might arise. Assigning unambiguous label names to process elements is a challenging task, particularly because process modelers are usually not experts in linguistics.

This paper presents algorithms that revise the vocabulary of process model element labels, which should increase the comprehension of the business process model. The algorithms are founded on effects from word recognition, which we applied to business process models, and also empirical results studying vocabulary preferences of process element labels. Figure 1 gives an overview of our approach. Exemplarily, the vocabulary revision algorithms should be applied on the business process model “handle exam results” (see the **input** business process

J. Zdravkovic et al. (Eds.): CAiSE 2015, LNCS 9097, pp. 69–83, 2015.
DOI: 10.1007/978-3-319-19069-3_5

model). Initially, the labels are extracted and segmented according to their part-of-speech (e.g., noun, verb) using a tagger. Additionally, structural information of the process element label[1] is stored (i.e., the position, the predecessor(s) and successor(s) of the label)[2]. The tagger also derives morphosyntactic information of labels (i.e., case, genus). After tagging, the labels are checked with respect to their linguistic fitness based upon a dictionary and/or on a domain ontology. The results of this analysis are linguistically revised process activity labels. Also no cleansing of the vocabulary might be required (if no indication for improvement is given) and the original label remains unchanged. The algorithms presented in this paper are based on the following methodological foundation. To understand how the vocabulary of labels might be improved, we checked effects from word recognition, which give hints how to efficiently recognize written words. These effects are explained in Section 2. Linguists are well-versed in applying these effects and therefore, we asked linguistics students to revise labels for exemplary business process models, which was part of a first empirical study. To benchmark the linguistic suggestions and to see whether users prefer the linguistically revised vocabulary, we performed a second empirical study with business process modelers. Both studies are summarized in Section 3. The algorithms for vocabulary revision are described in Section 4. The implementation of our approach is given in Section 5. Implications and limitations are discussed in Section 6. The



noun	verb	preset	postset
exam results	identify	view exam material	conduct post-exam review provide preliminary certificate

Fig. 1. Process of revising the vocabulary of process element labels

[1] The revision algorithms works on a process activity graph of business process models. The output of activities (e.g., places, events) are not considered since they use the same vocabulary but with a different grammatical conjugation.

[2] Note that the information of one activity label is stored within one row in order to keep the linguistic information.

add-on of our contribution is compared with related approaches in Section 7. Finally, the paper ends with a summary and an outlook.

2 Effects in Word Recognition

Process element labels are a concatenation of words. Frequently, a verb-noun style (e.g., initiate registration) is used to label process model elements. Common labeling styles are also a deverbalized-noun +"of"+ noun (e.g., evaluation of flights), a noun + deverbalized-noun (flight evaluation) or a gerund + noun (e.g., evaluating flights). Additionally, descriptives (e.g., by officer) or further part of speech (e.g., adjectives) can be used to label the process element[3]. Several effects have been identified in word recognition, which impact the access to words:

- *word frequency effect*: more often used (common) words are recognized more quickly in a text than less common words [3].
- *neighborhood frequency effect*: words are processed more slowly (errors can occur) when the neighboring words are orthographically similar to the stimulus word [4]. A further impact on this effect was found, which confirms that the existence of higher frequency neighbors facilitates processing of the stimulus word [5].
- *neighborhood size effect*: large neighborhood facilitates access for low-frequency words [5].

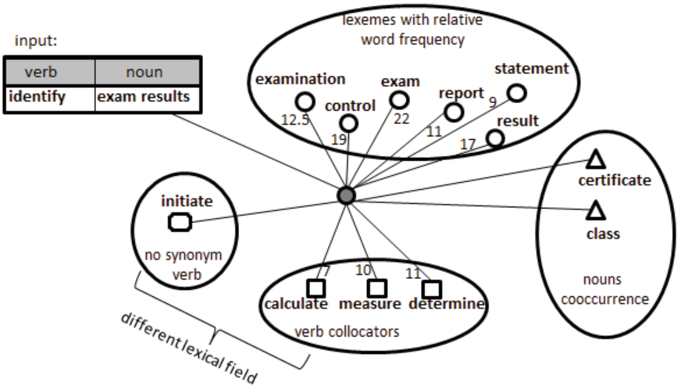
The word frequency effect is applied to process activity labels through assessing the *relative word frequency* of lexemes[4] of the original noun in a text corpus, a glossary or a domain ontology. The word frequency effect is also considered by determining an appropriate *collocation* of verbs and nouns[5]. Neighborhood frequency and size effects are addressed by the revision of the neighborhood of process activity labels. Each process activity has a direct dir_N and indirect ind_N neighborhood. Process activities, which directly precede (preset) or directly succeed (postset) a process activity are part of the direct neighborhood. Indirect neighbors are all remaining process elements of the business process model. The corresponding linguistic concepts that inspect the neighborhood are *co-occurrence* and *lexical field*. We consider co-occurrence as cohesion quality that is determined by mathematical and statistical computation and the results are to be interpreted. Lexical field is useful for the inspection of distinct activities of direct neighbors of a process activity label. Figure 2 applies exemplarily for the label "identify exam results" the four linguistic concepts. The term "exam

[3] The algorithms presented in Section 4 are implemented and tested for labels in German language. Terms discussed in this paper were translated to English by ourselves. The respective implications are discussed in Section 6.

[4] Particularly, we consider lexemes, which are in a synonym or hierarchical relationship (i.e., hypernyms, hyponyms), and lexemes, which belong to the same part of speech and have at least one common feature and thus belong to the same lexical field.

[5] Collocations are combinations of words that are preferred over other combinations that otherwise appear to be semantically equivalent [6].

result" is a compound, which however does not occur as one single term in the text corpus. Therefore, it is segmented into two terms, which are recognized as nouns by the tagger. Lexemes of the noun "exam" are "examination" (with rel. frequency of 12.5), "control" (with rel. frequency of 19) and "exam" (with rel. frequency of 22). Lexemes of the noun "result" are "report" (with rel. frequency of 11), "statement" (with rel. frequency of 9) and "result" (with rel. frequency of 17). Assume that terms with the highest word frequency are further considered. This means that "exam result" does not require any revision and remains "exam result" (both terms have the highest rel. word frequency among their lexemes). Verbs that occur near the term "exam results"[6] and are of the same lexical field than the verb "identify" are "calculate" (with rel. frequency of 7), measure (with rel. frequency of 10) and "determine" (with rel. frequency of 11). The collocator with the highest word frequency is "determine", which means that the original verb "identify" is replaced by "determine". For instance, a verb, which is not considered as collocator, is "initiate" because the verb does not belong to the same lexical field (i.e., is not a synonym of the input verb). Frequently the terms "certificate" and "class" cooccur with the term "exam", which means, it is checked whether exactly these terms are used in the business process model. If their synonyms were found then a replacement is performed (replacement of synonyms by the proper terms from the co-occurrence analysis). In the context


input:

verb	noun
identify	exam results

Fig. 2. Concepts used for revision of vocabulary

of an empirical study we asked students of German linguistics to revise process activity labels bearing in mind the four linguistic concepts. The intention of the study was to inspect the validity of the linguistic concepts. Subsequently, process modelers bench-marked the revised process activity labels. The results of the studies are presented in the next section.

[6] In case of compounds appropriate collocators are searched first for the second noun since the second noun is the primary word that refines the first noun.

3 Results from a Two-Stage Study

In the first stage of the study (run in December 2013), Bachelor students of German linguistics at the University of Greifswald had to revise the process activity labels of two business process models. The background qualifying the students to take part in the study is their attendance of a seminar of corpus linguistics. The first process model to revise was "handle and review exams", which should have been familiar to the students since all of them already should have passed through an exam. The second process model was the ITIL process "Incident Management", which should have been unfamiliar to the students. The process models were not designed with a particular modeling language. The students received an introduction to business process models and both process models were explained. Participants were free to answer the questions. The motivation to answer the questionnaire was a learning effect for the exam. To measure the understandability of the original process activity labels we used the perceived ease of use (PEOU) measure [7]. Prior to revision, the students were asked to complete on a Likert 5 point scale to each business process model the statement "It was easy for me to understand the meaning of the business process model".

Finally, we received 44 questionnaires, collected the suggestions in a spreadsheet file and applied a sorting for the suggestions. The revision suggestions were sorted according to identical names and identical labeling style. We observed that the participants mainly followed a verb-noun style (e.g., mark exam), which was also the predominant labeling style of the original process models. Also a deverbalized-noun+"of"+noun style (e.g., registration of students) was used. We observed that the deverbalized-noun+"of"+noun labeling style was used if the verb of the original activity label was unusual and instead a synonym noun was found to be more common (e.g., prevent problem → analysis of problems)[7]. Due to numerous domain-specific expressions in the ITIL process model, the students had difficulties (compared to the improvement of the "handle and review exams" process) to revise the labels. The suggestions for labels of the ITIL process model were less effective (i.e., we received a lower number of revision suggestions than for the other more understandable process model). The difficulty of revising the process activity labels of the ITIL process model is also indicated in the degree of the understandability measure. The cumulative frequency for PEOU of the original "handle and review exams" process model (judged by the students) is 38.46% for strong agree and agree, 38.46% for neutral and 23.1% for disagree and strong disagree. The cumulative frequency for PEOU of the original ITIL process model is 14.28% for strong agree and agree, 42.85% for neutral and 42.86% for disagree and strong disagree. These values indicate that more participants found the original "handle and review exams" process model easier to understand. For the ITIL process model it was vice versa.

In the second phase of the empirical study (run in February till April 2014) 49 process modeling beginners (graduates) and experts from different European

[7] An analysis in a German text corpus also indicates a higher relative frequency for the noun "analysis" ("Analyse") vs. "prevent" ("vorbeugen").

universities and research-driven institutes bench-marked the label suggestions in a paper-based questionnaire. The questionnaire was splitted up in group A and group B[8] and the participants had to indicate for each process model variant (original and revised labels) "It was easy for me to understand the meaning of the process elements" and to give a preference for a process model variant. The split into two groups should avoid crossover effects. The PEOU measures for the revised "handle and review exams" process model for group A are 87.2% and for group B 100.0% for strong agree and agree. The PEOU measures for the revised ITIL process model for group A is 77.9% and for group B 91.7% for strong agree and agree. These results mean that the revised process models were highly understandable for the respondents. To determine the degree of agreement (consensus) among interviewees of group A (beginners and experts) we used the Cohens Kappa coefficient. The coefficient has the value of 0.71, which indicates a good agreement among the interviewees. Thus, the understandability of process activity labels does not depend on modeling experiences. After reviewing the original process model versus the revised process model, the participants judged the usefulness of the process models against each other. Group A received first the original process model followed by the revised process model. Group B received the process models in a reversed order. Table 1 shows the statistical results for group A (left hand side) and Group B (right hand side). The process modelers preferred the revised "handle and review exams" process model over the original process model while the original ITIL process model was preferred over the revised ITIL process model by both groups. The conclusion from the two-stage study is that the application of effects from word recognition impacts the understandability and users also prefer linguistically superior labels. This assumption was observed for process models where no domain specific vocabulary was used. Therefore, when a common vocabulary is used it might be sufficient to access standard language dictionaries in order to revise the vocabulary. For domain dependent labels a domain ontology or a glossary is recommended in order to provide better suggestions for the revision of the vocabulary.

Table 1. Statistical test results for the process modeler preferences

Usefulness	$Mean_A$	St.Dev.	t-value	p	$Mean_B$	St.Dev.	t-value	p
$exam_o$ vs. $exam_r$	43.46	27.27	9.9526	< 0.0001	44.54	20.88	7.9815	< 0.0001
$exam_r$ vs. $exam_o$	69.23	26.32	16.4263	< 0.0001	67.27	24.52	10.2651	< 0.0001
$ITIL_o$ vs. $ITIL_r$	80.0	16.32	30.6127	< 0.0001	71.0	26.69	9.9535	< 0.0001
$ITIL_r$ vs. $ITIL_o$	38.18	17.99	13.2537	< 0.0001	50.96	25.0	7.627	< 0.0001

4 Revision Algorithm for the Vocabulary of Labels

Based on these findings from the study revision algorithms were designed.

[8] 37 persons (19 beginners, 18 advanced) answered the group A questionnaire. 12 persons (12 advanced) answered the group B questionnaire.

4.1 Preliminary Steps

Before the revision can be initiated, three preliminary steps are necessary. Firstly, a part-of-speech tagging (POS tagging) must be applied in order to assign parts of speech (e.g., noun, verb) to each term in the label [8]. The tagger should work on a data and tag set, which consider peculiarities of the language in use[9]. The part-of-speech assignment allows to categorize labels to a labeling style (i.e., verb-noun style, deverbalized-noun +"of"+ noun, noun + deverbalized-noun, gerund + noun). Secondly, process activities with labels using one or more composition operators (i.e. here: *and*, *or*) are decomposed into several atomic activities. The decomposition for the composition operator *and* is illustrated in Figure 3. The same holds for the decomposition of a process activity whose label contains the composition *or*. Such labels are decomposed into two (or more) process activities with the routing element XOR or possibly AND (this is case dependent). Thirdly, it is required to determine the subject area of the business process model. This step prevents that synonyms of higher relative word frequency but different subject area (e.g., a synonym of the term "exam" with higher word frequency is "monastery", which however does not fit in the context of exam) are considered as appropriate candidates. The algorithm determining the subject area is as follows (see Algorithm 1). All nouns of the business process model are extracted and the hypernyms for each noun are determined. The subject area corresponds to the most frequently found hypernym(s). Finally, all lexemes of the most frequently given hypernym(s) are extracted (this step is required for Algorithm 2).

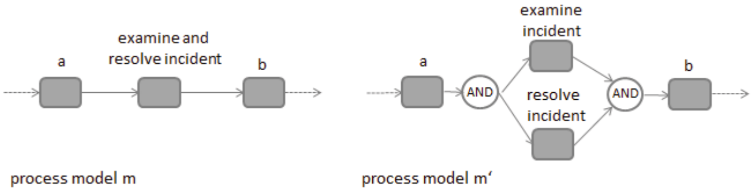


Fig. 3. Excerpt of a process model m before (left) and after preprocessing transformation into model m' (right)

4.2 Revision Algorithms

Based upon these preliminary steps, revision algorithms depending on the labeling style are executed (see Algorithm 2). The process of each algorithm is to analyze the noun(s) of the label (Step 1), subsequently the verbs undergo an analysis (Step 2). Finally, the vocabulary of the neighborhood of process activities is inspected, which might result in a further revision of the vocabulary of the label (Step 3).

[9] For instance, compounds are in the German language composite terms, which must be segmented to single terms by the tagger.

Algorithm 1. Algorithm to determine the subject area

```
input: ProcessModel model;
output: List subjectArea, List lexemes;
List elements = model.extractElements(); CountList hypernyms;
for all element: elements do
    for all noun: element.getNouns() do
        for all hypernym: noun.getHypernyms() do
            hypernyms.add(hypernym);
        end for
    end for
end for
List subjectArea; List lexemes;
for all hypernym: hypernym do
    if hypernym.count() $>$ 1 then
        subjectArea.add(hypernym);
        lexemes.add(hypernym.getLexemes());
    end if
end for
```

Step 1 of the vocabulary revision is identical for the four labeling styles, which is to determine the lexemes of each noun[10] with its relative word frequency. The algorithm extracts lexemes, which belong to same part of speech and have at least one common feature (e.g., synonyms or hierarchical relationship). A combination frequency is checked for compounds using co-occurrence (i.e., which lexemes of a compound are often combined with each other). Next, intersections between the extracted lexemes of each noun and the lexemes of the subject area (hypernym(s)) are determined. Lexemes of intersection with high relative word frequency are considered as potential candidates. Lexemes, which do not intersect the subject area and are part of compounds, are selected based upon the relative word frequency. All candidates are collected within a list.

Step 2 of the revision algorithm depends on the labeling style. Given a verb-noun style (see Algorithm 2), verb collocators are determined. It is searched for verbs, which are often combined with the candidate noun(s) and which are in the lexical field (e.g., synonyms) of the original verb as well. The collocator with the highest relative word frequency is considered as candidate. As subsequent step the algorithm checks if the verb candidates perform a distinct action to its neighborhood process activities. Particularly, the direct neighborhood dir_N is considered (see Section 2). When revising the vocabulary of process activities, it should be taken into account that process activities in the preset and postset should perform distinct actions. The lexical field theory [9] is applied for this purpose. For each process activity it is inspected if nouns of process activities in the direct neighborhood are synonyms or belong to identical lexical field. If so, then the verb collocator must be of different lexical field. Note that the structural

[10] All nouns of the element labels and compounds were already extracted and segmented into single terms (see preliminary steps).

transformation shown in Figure 3 does not affect the vocabulary analysis in terms of dir_N as $b_m\bullet = a$ and $b_{m'}\bullet = A$ and $\bullet c_m = a$ and $\bullet c_{m'} = A$ where $A = \{a_1, ..., a_n\}$[11]. An indirect neighborhood ind_N of a process activity includes all process activities that precede and succeed a process activity without its direct neighborhood. ind_N is used to revise the process activity label with respect to the linguistic concept of co-occurrence. For each original noun a co-occurrence analysis is performed based upon a matching of terms of occurrence.

Algorithm 2. Algorithm to revise the vocabulary of verb-noun style labels

```
input: ProcessElement element;
output: ProcessElement elementRevised;
List $dir_{Ns} = element.getDir_{Ns}()$;
List collocators = element.getVerb().getCollocators();
Collocator candidate = collocators.selectBestCandidate();
for all $dir_N : dir_{Ns}$ do
    if
    (isSynonym($dir_N$.getVerb(),element.getVerb()) OR
    identicalLexicalField($dir_N$.getVerb(),element.getVerb()) AND
    differentLexicalField($dir_N$.getVerb(),candidate) then
        elementRevised = candidate;
    else
        elementRevised = element;
    end if
end for
```

The algorithm to revise the vocabulary of a deverbalized-noun+of+noun style works as described in the following[12]. Here, the collocator is a deverbalized noun instead of a verb. After extracting the lexemes of the noun, lexemes of the deverbalized noun are determined. Subsequently, the combination of deverbalized noun + revised noun is compared versus revised noun + verb collocators of the same lexical field as the deverbalized noun (based upon rel. word frequency). The combination with the highest relative word frequency is selected. This algorithm is also applied for a noun+deverbalized-noun and a gerund+noun labeling style. The next section applies the algorithms for the input business process model in Figure 1.

4.3 Application of the Algorithms

Consider the input business process model in Figure 1. Initially, a part-of-speech tagging is applied and the labeling style "verb-noun" has been identified 10 or 11

[11] $\bullet x$ is called *preset*, which is the set of all preceding activities. $x\bullet$ is called *postset*, which is the set of all succeeding activities.

[12] Due to space restrictions we do not include pseudo code for this algorithm in this paper. However we plan to publish both source code and used data set online in the near future, in order to support the repeatability of our scientific work.

times respectively. The process activity “organize and check post-exam review” is decomposed into the process activities “organize post-exam review” and “check post-exam review” (after decomposition the verb-noun labeling style is used 11 times). The parts of speech are stored within rows of a table additionally with the labels of dir_N. Nouns of the considered business process model are “exam material”, “exam review”, “exam result”, “certificate”, “exam statistics”, “exam assessment”, “exam removal” and “process”. The subject area is determined by the most frequently used noun in this list, which is “exam”. Subsequently, hypernyms of the term “exam” are determined. Next, lexemes (hypernyms, hyponyms and synonyms) of the four hypernyms of “exam” are extracted. The results are shown in Table 2. Hypernyms and lexemes of the subject area are used when intersections between the lexemes of each original noun and the lexemes of the subject area (hypernym(s)) are determined. Prior to this step it is required to extract the lexemes of each original noun with its relative word frequency. Table 3 shows this process for an excerpt of nouns. In case of compounds it is checked whether and which lexemes frequently co-occur. For instance, the terms “exam” and “record” are more frequently combined than “exam” and “material”. The new noun of a label is the term, which is most frequently used and is in the range of the subject area. Lexemes of the terms “material” and “certificate” do not intersect with lexemes of the subject area. Therefore, a selection is done based upon the relative word frequency.

Table 2. Hypernyms and lexemes of the subject area term

most frequently used noun (subject area)	**hypernyms**	**lexemes**
exam	control, performance test, test, examination	control, performance test, test, examination, written test, exam, examination, exercise, inquiry, evaluation, study, assessment, activity, process.

Table 3. Suggestions for input nouns of process model in Figure 1

$noun_{orig}$	*lexemes*	comparison	$l_{intersec}$	$noun_{new}$
(exam) material	material (7.2), records (9.0), documentation (11.1), evidence (4.5), proof (3.5), paper (9.7)	exam - material (7.2), exam - records (11.2), exam - documentation (10.7)	-	(exam) record
certificate	certificate (10.0), testimonial (5.5), letter of reference (4.3), credentials (2.7), attestation (9.3), report (10.2)	-	-	report

Subsequently, verb collocators are determined. Exemplarily, verb collocators for the term “certificate” (the label is “provide preliminary certificate”), which are in the lexical field of the verb “provide” are “issue” (with rel. word frequency of 14.2), “acquire” (with rel. word frequency of 6.2), “prepare” (with rel. word frequency of 10.1) and “provide” (with rel. word frequency of 2.1). Since “issue” has the highest relative word frequency, it is considered as a better verb collocator than “provide”. After improving the collocators for all labels, the verbs of dir_N of a label are inspected. Label of dir_N of the label “issue preliminary certificate” are “determine exam results” and “end process”. Since the nouns “certificate”, “exam results” and “process” are not synonyms of each other, no further consideration is required. Finally, a co-occurrence analysis is performed. Figure 4 shows individual co-occurrence graphs for the terms “exam” and “result” to visualize terms of occurrence.

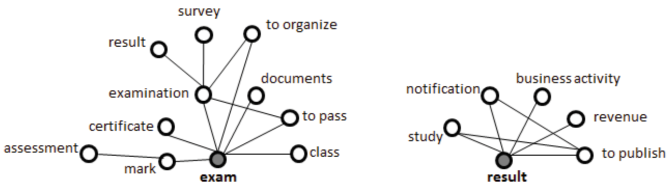


Fig. 4. co-occurrence graphs

Nouns frequently used with the term “exam” are “certificate”, “class”, “document”, “examination” and “mark”. The revision algorithms suggested to replace the term “certificate” by “report” (due to the higher rel. word frequency). Since the term “certificate” is more frequently used with the term “exam”, the term “certificate” is overwritten and is the candidate term.

5 Implementation and Analysis

The revision of the vocabulary of process activity labels has been implemented for the German language and a verb-noun labeling style. However, we do not expect that significant extensions are required to make the algorithms suitable for the English language. In contrast, the high frequency of morphological compounding in the German language makes finding of unknown words more difficult than for the English language [10].

To tag the labels the Stanford POS-Tagger is used. This tagger has been selected due to its high accuracy and its validation on the English and German language [11]. The tagger uses the negra corpus[13] and the STTS tag set[14]. Additionally, a list of German prefix and particle verbs has been created and

[13] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html

[14] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

is applied. Two German corpora are used to revise the vocabulary, which are DWDS[15] and COSMAS II[16]. The subject area of a business process model is derived from the component *OpenThesaurus* of DWDS. This thesaurus suggests synonyms, hypernyms and hyponyms of a query term. Verb collocators are derived through the *word field analysis* component of DWDS. The co-occurrence analysis is performed with COSMAS II. In the current implementation the revision based upon the co-occurrence analysis is done manually by the user. The tool suggests a list of nouns, which might be more appropriate than the revision. Thus, the revision of labels according to the co-occurrence concept depends on the decision of users. In future we plan to integrate machine learning techniques in order to make automatic suggestions based upon user's preferences.

The implementation has been validated on several business process models. We observed that the results clearly depend on the domain specificity of the vocabulary. We manually created a glossary for a set of ITIL process models and observed that both text corpora DWDS and COSMAS II were not suitable in this context. In such a context a domain ontology must be used. The initial analysis results underpin the results from the two-stage study where students of the linguistics acted according to a text corpus like DWDS and the revised process models were preferred over original process models for process models with common terms.

6 Discussion

Implications. From our point of view the revision approach has implications on all approaches that deal with process element labels. For instance, our approach concerns approaches detecting the similarity between process models. Algorithm 1, which determines the subject area of business process models can be used to uncover the semantic field between terms (similarity algorithms searches for semantic fields). Additionally, the revision algorithm might serve as a data cleaning approach before applying any approach for similarity calculation. It searches for similar terms in a context where lexemes of terms were identified and already revised. It is expected that similarity searches are performed more effectively and efficiently. Since linguistically revised process element labels strikes agreement, such a feature should also be integral part of a process modeling tool. This would relieve the process modeler from this manual task.

Limitations. The participants of the first study were linguistic students, which, from our point of view, qualified them revising process activity labels due to their background knowledge in corpus linguistics. However, the involvement of students always raises discussions about the external validity of the results. Although the students were not highly familiar with the process model paradigm (they received an introduction to business process modeling) the number of the highly qualitative suggestions for new process activity labels must be pointed

[15] http://www.dwds.de/

[16] www.ids-mannheim.de/cosmas2/

out. The quality was attested by an expert of the linguistics. Thus negative consequences due to the education level of the participants were not observed.

The techniques presented in this paper have been developed for the German language. The high morphological occurrence of terms in the German language makes the application of the revision approach even more difficult than for the English language. For instance, the English language has only few prefixes, which are a common feature in the German language[17]. The frequent usage of compounds in the German language also makes the revision approach more difficult than for the less morphological occurrence of compounds in the English language.

Lastly, the implementation of the revision algorithms is limited by ongoing research in corpus linguistics (e.g., the algorithms to detect homonyms are still not satisfying), which however, point to open research directions that must be tackled in order to improve the quality of business process models. Thus, the topic addressed in this paper also paves the way for additional research.

7 Related Work

Revising the vocabulary of process activity labels impacts the quality of the process element labels and finally of the complete business process model. Related approaches which also address the improvement of process element label quality perform this task by (1) improving the labeling style, or (2) assisting in the labeling of process elements.

Process model elements can be labeled according to several styles, which impact the understandability of the user in a different way. An empirical study of [12] found out that a verb-noun style is the preferable labeling style. Transformation algorithms exist, which convert an improper labeling style to this preferable style [13]. Although the preference for a verb-noun style over, e.g., a deverbalized-noun+“of”+noun style is comprehensible, no vocabulary revision has been performed for the process models used in this study. Our observation in the empirical study summarized in Section 3 was, that the preference for a labeling style highly depends on the familiarity of terms and thus a general recommendation for a verb-noun style might not be maintained. This observation calls for a further empirical study investigating preferences for labeling styles after vocabulary revision.

The second stream of related approaches automates the labeling of process model elements and thus relieves the manual and error-prone task, which is called to decrease label quality. Process model elements might be automatically generated based upon a glossary [14] or the linguistic analysis of process model elements [15]. Process element names are generated in the approach of [14] by a label suggestion component that also incorporates a label checker. The suggestion component works on a glossary being aware of control flow aspects of process models (this also allows to detect control-flow errors during labeling).

[17] http://www.bu.edu/isle/files/2012/01/-Stefan-Diemer-Corpus-Linguistics-with-Google.pdf

The glossary is created from a given collection of process models without improving the vocabulary. The approach of [15] makes suggestions for labels based upon the inspection of labels that were gathered from a collection of business process models. Particularly, suggestions for element labels are based on the analysis of holonyms (a word representing the whole of a part-of relation) and hypernyms (a more general word) relationships. The authors propagate their approach for finding element labels for process model abstraction. This is also reasonable due to the limitation of analyzing holonyms and hypernyms. The label repository gathered by [15] might be suitable as foundation for revision techniques suggested in this paper. For instance, relative word frequency count and verb collocation might be used to determine the label with the highest linguistics among all similar labels. Thus, synergies can be found here. A further approach related to the assistance of element labeling is suggested by [16]. This approach detects naming conflicts already during the modeling process using a repository of domain specific vocabulary. The approach of [16] might be complementary when a business process model is described by a domain specific vocabulary. For instance, a glossary or domain ontology might be created from the domain specific repository. A domain ontology supports the revision of domain specific process models.

To sum up, our approach can be considered as a preprocessing step for most of these related approaches that could profit from an adjusted and improved vocabulary since these approaches rely on the labeling of process model elements.

8 Conclusion and Outlook

The revision of the vocabulary of process activity labels is connected with the quality of business process models. Several approaches addressed the quality improvement of process element labels by, for instance, postulating a labeling style, which improves comprehension. This paper suggested the revision of the vocabulary of process element labels as a research step to improve the quality of business process models. Four linguistic concepts were applied to business process elements. These concepts were derived from word recognition effects that give hints how to better recognize written words. Results from an empirical study indicate the validity of these concepts. In the future we plan to incorporate the analysis of all common labeling styles and to perform the analysis for the English language. To finally find determinants of the understandability of process element labels we are also conducting several empirical studies investigating the visual design of element labels (e.g., their textual segmentation).

References

1. Wilmont, I., Brinkkemper, S., van de Weerd, I., Hoppenbrouwers, S.: Exploring intuitive modelling behaviour. In: Bider, I., Halpin, T., Krogstie, J., Nurcan, S., Proper, E., Schmidt, R., Ukor, R. (eds.) BPMDS 2010 and EMMSAD 2010. LNBIP, vol. 50, pp. 301–313. Springer, Heidelberg (2010)

2. Moody, D.: The ‘physics’ of notations: Toward a scientific basis for constructing visual notations in software engineering. IEEE Trans. Softw. Eng. **35**, 756–779 (2009)
3. Balota, D.A., Spieler, D.H.: Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. Journal of Experimental Psychology: General **128**, 32–55 (1998)
4. Grainger, J., ORegan, J.K., Jacobs, A.M., Segui, J.: On the role of competing word units in visual word recognition: the neighborhood frequency effect. Percept Psychophys **45**, 189–195 (1989)
5. Sears, C.R., Hino, Y., Lupker, S.J.: Neighborhood size and neighborhood frequency effects in word recognition. Journal of Experimental Psychology: Human Perception and Performance **21**, 876–900 (1995)
6. Croft, W., Cruse, D.A. (eds.): Cognitive Linguistics. Cambridge University Press (2004)
7. Maes, A., Poels, G.: Evaluating quality of conceptual modelling scripts based on user perceptions. Data Knowl. Eng. **63**, 701–724 (2007)
8. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Brill, E., Church, K. (eds.): Proceedings of the Empirical Methods in Natural Language Processing, pp. 133–142 (1996)
9. Coseriu, E., Geckeler, H.: Trends in Structural Semantics. Tübinger Beiträge zur Linguistik, Narr (1981)
10. Tseng, H., Jurafsky, D., Manning, C.: Morphological features help pos tagging of unknown words across language varieties. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp. 32–39 (2005)
11. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP 2000. Association for Computational Linguistics, pp. 63–70 (2000)
12. Mendling, J., Reijers, H.A., Recker, J.: Activity labeling in process modeling: Empirical insights and recommendations. Inf. Syst. **35**, 467–482 (2010)
13. Leopold, H., Smirnov, S., Mendling, J.: On the refactoring of activity labels in business process models. Information Systems **37**, 443–459 (2012)
14. Peters, N., Weidlich, M.: Using glossaries to enhance the label quality in business process models. In: Proceedings of the 8th GI-Workshop Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten (EPK), CEUR-WS.org, pp. 75–90 (2009)
15. Leopold, H., Mendling, J., Reijers, H.A., Rosa, M.L.: Simplifying process model abstraction: Techniques for generating model names. Information Systems **39**, 134–151 (2014)
16. Delfmann, P., Herwig, S., Lis, L., Stein, A.: Supporting distributed conceptual modelling through naming conventions - a tool-based linguistic approach. Enterprise Modelling and Information Systems Architectures **4**, 3–19 (2009)