Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4606–4610
Marseille, 20-25 June 2022

German Light Verb Constructions in Business Process Models

Kristin Kutzner, Ralf Laue
University of Applied Sciences Zwickau, Germany
Dr.-Friedrichs-Ring 2a, 08056 Zwickau
Kristin.Kutzner@fh-zwickau.de, Ralf.Laue@fh-zwickau.de

Abstract

We present a resource of German light verb constructions extracted from textual labels in graphical business process models. Those models depict the activities in processes in an organization in a semi-formal way. From a large range of sources, we compiled a repository of 2,301 business process models. Their textual labels (altogether 52,963 labels) were analyzed. This produced a list of 5,246 occurrences of 846 light verb constructions. We found that the light verb constructions that occur in business process models differ from light verb constructions that have been analyzed in other texts. Hence, we conclude that texts in graphical business process models represent a specific type of texts that is worth to be studied on its own. We think that our work is a step towards better automatic analysis of business process models because understanding the actual meaning of activity labels is a prerequisite for detecting certain types of modelling problems.

Keywords: German, light verb construction, support verb construction, Funktionsverbgefüge, business process, business process model

1. Background of our Research

1.1. Business Process Models

Business process models (BPM) are visual representations of processes in an organization. They can serve as a base for communication between the stakeholders in a process improvement project. They can also be used for training new employees or for verifying that legal compliance rules are enforced. Last but not least, BPM are used for automating processes using software applications that manage to keep track of the state of the process, to distribute work and related data and to execute certain steps (such as computations) automatically.

Usually, BPM are created using visual languages specifically designed for modelling business processes, such as BPMN (Object Management Group (OMG), 2011) or Event-Driven Process Chains (EPCs). As the majority of BPM analyzed in our work use the modelling language EPC, the specifics of textual labels will be discussed by example of EPC diagrams. EPCs consist of functions (tasks that need to be executed, depicted as rounded boxes), events (pre- and postconditions before/after a function is executed, depicted as hexagons) and connectors (which can split or join the flow of control between the elements, allowing to model parallel and alternative executions). Arcs between these elements represent the control flow. From the example in Fig. 1, it can be observed that a common way to label a function is to use a verb phrase containing a verb and its dependents (in particular, one or more objects). Also, it can be seen that events serve two purposes. They can describe a (pre-)condition for executing a task or a state that can be observed at some point of time in a process. Note that only very rarely the labels contain complete sentences.

To fulfill their purposes, it is imperative that BPM correctly describes the steps in the process flow and the order between them. As both experience as research has shown that far too often BPM contain errors (see e.g. (Mendling et al., 2006)), a lot of research has been conducted on preventing these errors. Based on this research, advanced modelling tools have been developed that can analyze the models automatically.

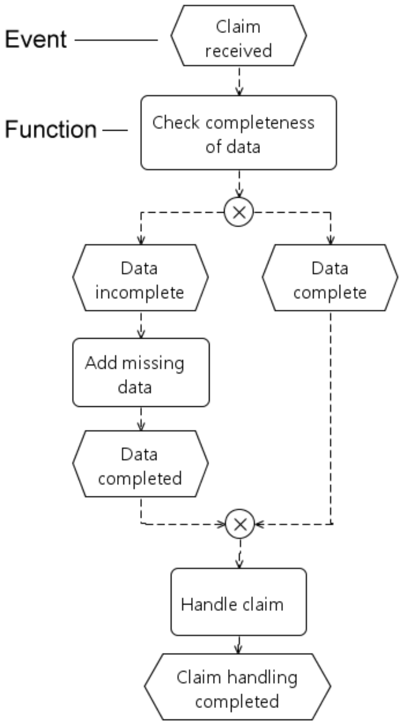


Figure 1: EPC model fragment

For such an analysis, two aspects need to be taken into account: First, the order of the graphical elements (usually expressed by boxes, arrows and other shapes) and second, the natural language labels that are attached to the graphical elements. While for the purpose of this work, the formal analysis of the graphical elements is out of scope, we will shortly discuss existing approaches for analyzing the textual labels. In (Laue et al., 2016), the textual descriptions of tasks in BPM are analyzed in order to detect several potential modelling problems such as:

- tasks that are described too vaguely,
- tasks for which it is likely that their execution is modelled in the wrong order (e.g. "send something" takes place before "pack something"),
- contradicting tasks that can be executed at the same time (e.g. "accept proposal" is executed as well as "deny proposal").

It is easy to see that for this kind of analysis, it is necessary to derive the type of activity (something as "to send", "to measure" or "to deny", i.e. something expressed by a verb) as well as its dependents (most importantly, an object that is expressed by a noun) from the task labels. A common modelling guideline suggests to label tasks in verb-object style, and there exist approaches such as (Leopold et al., 2013) to analyze models automatically and to make the modeller aware of task labels that do not conform to this style.

Working on the analysis of BPM in German, light verb constructions constitute a major challenge. For example, a naive analysis by means of POS tagging would come to the conclusion that if a function is labeled by the phrase *Einblick nehmen* (to gain insight), it would be the Verb *nehmen* (to take) that describes the type of activity while *Einblick* is the affected object. While this is grammatically correct, in fact no information at all is given about the affected object of the business process activity. This makes the analysis of BPM much more complicated – rightfully, (Sag et al., 2002) state that light verb constructions and other multiword expressions can be a "pain in the neck for NLP".

Given the fact that such constructions appear quite frequently in German business process models (more detailed statements about the frequency will follow below), we had reason to study the light verb constructions in German BPM in order to analyze those models.

1.2. Light Verb Constructions

There is no generally accepted definition of the term "light verb construction" (LVC) or the term "Funktionsverbgefüge" in German (van Pottelberge, 2008; Harm, 2021). Existing definitions have in common that they refer to a collocation where a verb – the light verb (also called function verb or support verb) – is a main verb that has lost most of its concrete lexical semantics. It is not mainly the verb but its collocate that describes an action. Some authors (such as (Krenn, 2008)) restrict the collocates that form a LVC together with the verb to predicative nouns which is also the usual definition of the German term "Funktionsverbgefüge" (see (van Pottelberge, 2008)). For the purpose of our paper, we use a wider definition and consider (preposition-determiner)-verb-noun collocations (*Bericht erstatten* = to report) as well as (preposition-determiner)-verb-adjective collocations (*geboten sein* = to be imperative). In addition, we also list phrases where a nominalized verb (to be more specific: *nomen actionis*) occurs together with words such as "execute", "do", "undertake" (in German e.g. *durchführen*, *ausführen*, *vollziehen*, *vornehmen*, etc.). Finally, our list contains phrases such as *Kopie anfertigen* (to make a copy). While this is a normal verb phrase consisting of verb and object, the object explains the action more specifically than the verb, and the phrase could be replaced by the verb *kopieren* (to copy).

We did, however, not include phrases such as *Prüfung beginnen* (to start an examination), i.e. cases where the verb indicates whether an activity starts/is started, an activity is carried on or an activity ends/is ended. These phrases do not describe the action itself (to start an examination is not equal to the verb to examine) and, therefore, should be considered separately.

2. Related Resources

(Krenn, 2008) provides a data set of preposition-noun-verb combinations that have been extracted from the Frankfurter Rundschau corpus. 549 of them have been classified (by means of manual annotation) as LVC and 600 as figurative expressions.

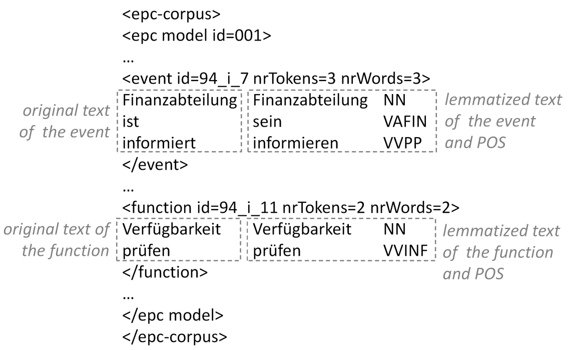
(Marušić, 2015; Marušić, 2018) analyzed 10 business reports from large German corporate groups. 7,327 occurrences of LVC were extracted from 37,982 sentences. An interesting observation from this study was that for 13.9% of those LVC, no synonymous verb exists that could replace the LVC. Furthermore, it was found that the frequency of LVC in this type of texts is substantially higher than frequencies known from texts in other corpora.

(Bruker, 2013) compiled a list of more than 2,000 German LVC from an extensive corpus analysis that included the TIGER and DWDS corpus.

(Kamber, 2008) used a corpus of 52 editions of the German news magazine "Der Spiegel" published in 1997 (5 mio words) and a corpus containing articles from the Swiss daily newspaper "Tages-Anzeiger" (1996-2000, 61 mio words) for extracting a list of the 10 most frequent LVC for each of the 10 support verbs *setzen, stellen, stehen, nehmen, bringen, kommen, geraten, gehen, sich befinden* and *bleiben*. Kamber's research is motivated by the desire to teach German light verb constructions to non-native speakers.

3. Research Design

To identify which LVC are used in BPM, we used a three-staged research design. To start with, we collected and extracted BPM (Stage 1) before preparing a model text corpus in Stage 2. Finally, based on the resulting corpus, we manually identified and catalogued LVC (Stage 3).



```
<epc-corpus>
<epc model id=001>
…
<event id=94_i_7 nrTokens=3 nrWords=3>
Finanzabteilung    Finanzabteilung   NN
ist                sein              VAFIN
informiert         informieren       VVPP
</event>
…
<function id=94_i_11 nrTokens=2 nrWords=2>
Verfügbarkeit      Verfügbarkeit     NN
prüfen             prüfen            VVINF
</function>
…
</epc model>
</epc-corpus>
```

original text of the event
lemmatized text of the event and POS
original text of the function
lemmatized text of the function and POS

Figure 2: Extraction of the prepared business process model corpus

Stage 1: Collecting and extracting business process models Firstly, we searched for BPM in all sources that were available to us. We decided to include models drawn in the modeling language Event-Driven Process Chain (EPC) (Scheer et al., 2005), a widely used language for representing business processes. Doing so, we selected 2,301 German BPM from public repositories, textbooks, scientific papers, student papers and real-world projects:

- 604 models from the SAP R/3 reference model (Keller and Detering, 1996), a widespread business reference model
- 393 models from the repository of the BPM Academic Initiative (Kunze et al., 2012)
- 349 models from 17 real-world projects
- 329 models from 26 textbooks
- 210 models from 12 bachelor, 3 master and 43 diploma theses and 10 term papers
- 106 models from 17 various resources found on the internet (not belonging to any of the other categories)
- 83 models from 43 published scientific papers
- 54 students' solutions to university exercises
- 52 models from 13 PhD theses
- 39 models from a German process sharing platform for public administration processes (Eid-Sabbagh et al., 2011)
- 38 models from the publicly available repository of the process modelling tool Oryx (Decker et al., 2008)
- 22 models from 11 university lecture notes
- 14 models from 3 technical manuals
- 6 models that come as examples with the process modelling tool ARIS Toolset
- 2 models that have been published as a challenge in a workshop dealing with computer-supported correction of BPM

The selected models were extracted from the sources. If the models were available in print only, a computer-readable file has been created using the modelling tool bflow* Toolbox[1]. Depending on the source, the models were available in different file formats (e.g. .epc, .xml, .epml), which were then converted into one uniform file format.

Stage 2: Preparing a text corpus To prepare a text corpus, we implemented Python scripts that used the Natural Language Toolkit (NLTK). As a first step, we checked whether the data set contained data extraction-related noise and cleaned the text (e.g. by applying the same character encoding for mutated vowels). Then, the textual labels of functions and events were extracted and split into tokens by using the NLTK tokenizer package. Based on the tokens, we performed lemmatization and part-of-speech (POS) tagging using the Hanover Tagger which supports annotations based both on heuristics and on hidden Markov models of German morphology (Wartena, 2019). In addition, we determined the number of tokens and words of each function and event. Subsequently, we prepared an annotated BPM corpus by listing the textual labels of functions and events with their respective linguistic annotations for each model (see Fig. 2).

Stage 3: Investigating the text corpus and identifying light verb constructions In stage 3, we generated a list of all verbs identified in the corpus, together with the text passages of functions and events in which the verbs occurred. We manually inspected the list to identify and collect LVC. This manual inspection was done independently by both authors of his paper. If the authors came to different conclusions on whether a phrase should be included, each case was discussed until consensus was reached. Doing so, we built the resource of German light verb constructions in business process models which consists of three columns: a light verb (column 1), an identified LVC (column 2) and a corresponding full verb if available (column 3). For example, the light verb *stellen* (to put) is assigned to the LVC *Antrag stellen* (to submit a request) and the corresponding full verb *beantragen* (to request)[2]. There are several cases where more than one full verb could be assigned, and we make no claim that our choice is always the best one. In total, we manually analyzed 30,384 text passages and identified 846 different LVC. Finally, we determined the frequency distribution of light verb constructions in our text corpus[3] and identified a total of 5,246 light verbs [4].

4. Quantitative Statistics

Altogether the 2,301 EPC models contained 31,867 events and 21,096 functions. This means that on average, an EPC has 13.85 events and 9.17 functions. The event labels contained on average 3.46 tokens, the function labels 3.15 tokens. The longest event label had 22

[1] http://www.bflow.org

[2] see file 01_light_verb_constructions_ordered.txt in our resource.

[3] see file 02_light_verb_constructions_frequencies.txt

[4] see file 03_light_verb_construction_instances_found_in_corpus.txt

events	31,867
functions	21,096
in the labels of those events and functions:	
tokens	183,611
words	176,821
among those tokens were:	
nouns (NN & NE)	77,583
verbs (V*)	57,178
adjectives (ADJA & ADJD)	12,934
pre- and postpositions (A*)	12,321
articles (ART)	5,704
particles (PTK*)	5,663
conjunctions (K*)	2,924
numbers (CARD)	2,318
pronouns (P*)	1,441
adverbs (ADV)	996
others	9,259

Table 1: Results of POS-tagging for event and function labels

Antrag stellen (to file an application)	159
Plan erstellen (to draw up a plan)	148
Planung durchführen (to plan)	129
Zeit festlegen (to set time)	77
Prüfung durchführen (to conduct an examination)	74
Bewertung durchführen (to assess)	63
aktenkundig machen (to record something)	60
Rechnung durchführen (to compute)	59
Nummer vergeben (to assign a number)	55
Prüfung ablegen (to take an examination)	51
Analyse durchführen (to carry out an analysis)	50

Table 2: LVC with at least 50 occurrences

tokens, the longest function label 21 tokens. The results of the POS tagging are shown in Tab. 1. The main aim of applying POS tagging was to locate the verbs. Manual inspection showed that the POS tagger worked well for this task. This deserves to be noted because usually the quality of POS taggers is tested for whole sentences only.

5. Observations

Tab. 2 contains LVC that have been found most frequently in our corpus. Comparing the LVC in Tab. 2 with the statistics from (Marušić, 2015) shows that the language in BPM clearly differs from those in financial reports, although both belongs to the language of business. There is no LVC that can be found both in Tab. 2 and in the list of the 38 most frequent LVC that have been found in the corpus of financial reports. The most frequent LVC in the list published in (Marušić, 2015) (*zur Verfügung stehen* = to be available) was found 15 times in our corpus of BPM. The other way around, the most frequent LVC in our corpus (*Antrag stellen* = to file an application) is not among the list of 38 most frequent LVC (those with at least 15 occurrences) from (Marušić, 2015).

We conclude that BPM constitute a specific type of text that is worth to be studied on its own. Please note however, that we cannot claim that the frequencies from Tab. 2 are representative for German BPM in general. This can be explained by the way the BPM in the corpus have been selected. For example, it was quite often the case that two similar BPM (e.g. a model of the current situation and one of a future improved process) have been published in the same text. As a consequence, both models (that shared common labels) were added to the repository. Even more repetitions can be found in the solutions to university exercises where several models described the same process.

6. Conclusion

In this paper, we have presented a resource containing light verb constructions that appear in German business process models, especially in EPCs. If possible, corresponding full verbs were added manually. To extend and verify the resource, further business process models can be analyzed. Considering non-German models or models created with other visual modelling languages such as BPMN might be valuable future steps. Overall, our work contributes to research on textual labels in business process models. It serves as a foundation for automatically analyzing these models and for inferring the meaning of their texts. For example, researchers are supported in identifying modelling problems such as inconsistently labeled activities in a process model (e.g. using different phrases such as “to copy”, “to make a copy”, “to create a copy” instead of a consistent phrase).

Ultimately, we would like to note that the language in BPM is worth to be studied for other purposes as well. The nature of BPM offers interesting possibilities to learn temporal relations between activities – something that is out of the scope of this paper. The reason is that those models contain both information expressed as formal modelling language (the “shapes and arrows”) as information expressed in natural language (the labels of the model elements). Considering both kinds of information, one can draw conclusions on which activities follow each other, which activities often occur in the same business case, are executed at the same time in parallel or exclude each other. Extracting and exploiting such information will be the subject of future work.

7. Acknowledgements

Our research project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 445156547.

8. Bibliographical References

Bruker, A. (2013). *Funktionsverbgefüge im Deutschen*. Bachelor + Master Publishing.

Decker, G., Overdick, H., and Weske, M. (2008). Oryx - an open modeling platform for the BPM community. In *Business Process Management, 6th International Conference, BPM 2008, Milan, Italy, September 2-4*, volume 5240 of *Lecture Notes in Computer Science*, pages 382–385. Springer.

Eid-Sabbagh, R., Kunze, M., and Weske, M. (2011). An open process model library. In *Business Process Management Workshops - BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011*, volume 100 of *Lecture Notes in Business Information Processing*, pages 26–38. Springer.

Harm, V. (2021). *Funktionsverbgefüge des Deutschen: Untersuchungen zu einer Kategorie zwischen Lexikon und Grammatik*. De Gruyter.

Kamber, A. (2008). *Funktionsverbgefüge - empirisch*. Max Niemeyer Verlag, Tübingen.

Keller, G. and Detering, S., (1996). *Process-Oriented Modeling and Analysis of Business Processes using the R/3 Reference Model*, pages 69–87. Springer US.

Krenn, B. (2008). Description of evaluation resource – German PP-verb data. In *Proceedings of the LREC2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 7–10, Marrakech, Morocco.

Kunze, M., Berger, P., and Weske, M. (2012). BPM academic initiative - fostering empirical research. In *Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012), Tallinn*, volume 940 of *CEUR Workshop Proceedings*, pages 1–5. CEUR-WS.org.

Laue, R., Koop, W., and Gruhn, V. (2016). Indicators for open issues in business process models. In *Requirements Engineering: Foundation for Software Quality - 22nd International Working Conference, REFSQ 2016, Gothenburg, Sweden, March 14-17*, volume 9619 of *Lecture Notes in Computer Science*, pages 102–116. Springer.

Leopold, H., Eid-Sabbagh, R., Mendling, J., Azevedo, L. G., and Baião, F. A. (2013). Detection of naming convention violations in process models for different languages. *Decis. Support Syst.*, 56:310–325.

Marušić, B. (2015). *Funktionsverbgefüge in deutscher Konzernsprache*. Phd thesis, University of Osijek, Croatia.

Marušić, B. (2018). Besonderheiten der Funktionsverbgefüge in der deutschen Konzernsprache. *Jezikoslovlje*, 19(1):87–106.

Mendling, J., Moser, M., Neumann, G., Verbeek, H. M. W., van Dongen, B. F., and van der Aalst, W. M. P. (2006). Faulty EPCs in the SAP reference model. In *Business Process Management, 4th International Conference, BPM 2006, Vienna, Austria, September 5-7, 2006*, volume 4102 of *Lecture Notes in Computer Science*, pages 451–457. Springer.

Object Management Group (OMG). (2011). Business process model and notation (BPMN) version 2.0. Technical report.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15.

Scheer, A., Thomas, O., and Adam, O. (2005). Process modeling using event-driven process chains. In Marlon Dumas, et al., editors, *Process-Aware Information Systems: Bridging People and Software Through Process Technology*, pages 119–145. Wiley.

van Pottelberge, J., (2008). *Funktionsverbgefüge und verwandte Erscheinungen, in: Phraseologie, Volume 1*, pages 436–444. De Gruyter Mouton.

Wartena, C. (2019). A probabilistic morphology model for German lemmatization. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.