Trends on the Usage of BPMN 2.0 from Publicly Available Repositories

Ivan Compagnucci[0000−0002−1991−0579], Flavio Corradini[0000−0001−6767−2184], Fabrizio Fornari[0000−0002−3620−1723], and Barbara Re[0000−0001−5374−2364]

University of Camerino, Via Madonna delle Carceri 7, Camerino - 62032, Italy,
{ivan.compagnucci,flavio.corradini,fabrizio.fornari,barbara.re}@unicam.it

Abstract. Business Process Model and Notation is the de facto standard for graphically modelling business processes. Since its first release in 2004, it evolved until reaching the actual 2.0 version, which presents more than 85 elements. Despite the notation being rich in graphical elements, initial studies show that only a subset of the BPMN elements is actually used. This paper aims at investigating whether the BPMN vocabulary adopted nowadays by model designers shows some particular trends. We collected 25,590 models from six online repositories to conduct such an investigation, and we analysed them. We report and discuss the obtained results providing insights on the correlations in the BPMN vocabulary and the resulting complexity of BPMN models.

Keywords: BPMN · Business Process Modelling · Models Repositories

1 Introduction

Business Process Model and Notation[1] (BPMN) is an OMG standard [19] which provides a graphical notation for the modelling of business processes. The notation emerged as the de facto standard to support business process modelling for all the business stakeholders (e.g., business analysts who create and refine the processes, technical developers who implement them, business managers who monitor and manage them). The success of BPMN comes from its versatility and capability to represent business processes for different purposes. The intuitive graphical representation of more than 85 elements made BPMN widely accepted by both the industry and the academia.[2] Thanks to its intuitiveness, BPMN can be easily used to design processes and their interactions. These models can be used to communicate and interchange the business requirements of a business process and provide the underpinning of the actual process implementation. Therefore, a BPMN model can be understood by all stakeholders involved in the process. Since its first release (May 2004), BPMN has been also protagonist of

[1] BPMN 2.0.2 is the latest version released in December 2013 https://www.omg.org/spec/BPMN/2.0.2 formally published by ISO as the 2013 edition standard: ISO/IEC 19510.

[2] More than 70 tools support BPMN (http://www.bpmn.org)

several research contributions ranging from studies related to: the usage of the notation (e.g., [1, 10, 18, 21, 3]), the definition of a rigorous semantics for each notation element (e.g, [6, 8, 16, 22]), the evaluation of BPMN models qualities (e.g., [5, 11, 12, 2]), the definition of notation extensions to incorporate specific application domain aspects (e.g., [4, 23]), and many others.

This paper reports the results of an investigation on whether the BPMN vocabulary adopted nowadays by model designers shows some particular trends. We collected a total of 25,590 models from six collections, some of which have already been used for conducting research activities. We focused on studying the overall usage of modelling elements and their correlation and combined usage in designing a business process model. We also investigated the complexity of the notation, highlighting the actual difference between the BPMN practical and the theoretical complexity. From the conducted study, we conclude whether or not the usage of the BPMN notation for modelling business processes has changed during the years, especially comparing our results with previous studies conducted on the topic [18].

The paper is organised as follows: Section 2 shows the methodology we used for harvesting BPMN models. Section 3 reports the overall usage of BPMN notation. Section 4 reports about model complexity in terms of model size and variety of elements used. Section 5 highlights the correlation in modelling BPMN elements in combination and also the set of most popular vocabulary subsets of BPMN. A comparison with related works is presented in Section 6. Finally, Section 7 discusses limitations and future work.

2 Models Harvesting

For conducting our analysis, we gathered different repositories of business process models designed with the BPMN notation. In particular, we refer to six repository such as: BIT process library[3], Camunda BPMN[4], eCH-BPM, GenMyModel, GitHub, and RePROSitory.

The *BIT process library*, is composed by 850 BPMN files containing abstract business process used by IBM WebSphere Business Modeler V6 and V7. Those models have been made available by IBM for the practical validation of the soundness-checking approaches and tools [9]. The *Camunda BPMN* collection stores a total amount of 3,739 unique BPMN models designed in BPMN which have been made available, on GitHub[5], by the CAMUNDA company. These diagrams have been designed in BPMN training sessions, which have been given since 2008 until 2015 when the models have been released. The *eCH-BPM* model collections includes 117 models downloaded from the eCH process platform[6]. The published process models are examples of Swiss public administration's processes

[3] The full collection name is "IBM Research GmbH, BIT process library, release 2009"

[4] The full collection name is "Camunda BPMN for Research"

[5] https://github.com/camunda/bpmn-for-research

[6] http://www.ech-bpm.ch/de/process-library

that, have been designed by the eCH association, in cooperation with municipalities, cantons and federal agencies as well as the BPM specialist community. The *GenMyModel* collection refers to 11,460 models that have been downloaded, at the time of writing, from the GenMyModel platform[7]. Users can experiment with the platform's functionalities provided that the designed models will be made publicly available for reuse. The *GitHub* collection of models actually consists of a set of 17,203 BPMN models publicly available on GitHub that have been retrieved by Heinze et al. with a procedure described in [14]. The *RePROSitory* collection consists of 560 models that have been harvested, from the proceedings of the BPM conference and manually re-designed with the objective of defining a benchmark of models with a solid literature background. The models are available on the RePROSitory[8] dedicated platform [7].

For each collection of models we run a filtering procedure for removing models with a total number of element less than eight. We derived this threshold value by analysing each models collection and manually inspecting all the models which size was included in a range of zero to ten elements. We concluded that models with a number of elements less than eight were incomplete models that could have compromised the validity of our study. Table 1 reports the amount of models considered from each collection after the application of the filtering procedure. At the end of our models harvesting procedure, we gathered a total of 25,590 models.

Models Repositories	**Number of Considered Models**	**Source**
BIT process library	804	Literature
Camunda BPMN	3 721	Training sessions
eCH-BPM	117	Government
GenMyModel	11 156	Mixed
GitHub	9 232	Mixed
RePROSitory	560	Literature

Table 1: Models collections overview

For extracting data from the harvested models, we developed a python script which allows to count the occurrences in a *.bpmn* file of 85 BPMN elements. We based the development of the python script on the tags present in the BPMN 2.0 meta-model[9] so to be able to select all the actual BPMN elements plus some of them characterised by the usage of attributes. We ran our script over the 25,590 retrieved models. The source code and all the details about the usage of this script are reported on the PROS Lab GitHub account[10] together with all the extracted data and the data resulting from the performed statistical analysis.

[7] https://www.genmymodel.com/

[8] http://pros.unicam.it/reprository

[9] https://www.omg.org/spec/BPMN/2.0/About-BPMN/

[10] https://github.com/PROSLab/BPMN-element-counter

3 Overall Usage of BPMN Elements

In this section, we present statistical analysis performed over the entire set of models to detect which BPMN elements have been used to design such models and their frequency.

3.1 Distribution of BPMN Elements over Models

Considering the frequency distribution of the individual BPMN elements over the total amount of models we ranked them from the most frequent element to the less frequent. As it can be observed in Fig. 1, the distribution of BPMN elements follows a power-law distribution. In our study we found out that five elements namely *Sequence Flow, End None Event, Start None Event, Task, and Exclusive Gateway* are present in **more than the 50%** of the overall models. In the range **between 50% and 20%** we found 8 elements respectively Pool, User Task, Lane, Parallel Gateway, Message Flow, Start Message Event, Association and Service Task. **Between 20% and 10%** we found 7 elements respectively *Intermediate Catch Timer Event, Intermediate Catch Message Event, End Terminate Event, Event Based Gateway, Collapsed Sub Process, Conditional Event, Text Annotation.* The remaining elements of the notation (65 of 85) are present in less than 10% of models. **Between 10% and 1%** we found 25 elements mainly related to typed tasks, some particular type of events, data related elements (*Data Object* and *Message*), grouping elements (*Group*), *Expanded Sub-Processes* and *Default Flow.* 36 elements over 85 are present in less than 1% of the models, they mainly include particular types of Activities (*Transaction*, *Ad Hoc Sub Processes*, *Task Loop Activity*, etc.), particular types of Events (*Intermediate Catch Multiple Event, Intermediate Throw Multiple Event, etc.*) and other elements such as *Choreography Task*, *Conversation*, etc. To deepen our analysis, we also derived and reported in Fig. 2 the average number of occurrences of a given construct in a model. The reported values are only related to models that present such a construct. From Fig. 1 we notice that 99.65% of the models include a Sequence Flow (ID 1),[11], while from Fig. 2 we notice such a percentage of models presenting around 16 sequence flows on average. Some peaks can be seen in Fig. 2 even in correspondence of elements that are not present in many models. It is the case of Choreography Participant (ID 72). It is widespread in 0.07% of the models.

3.2 Frequency Distribution of BPMN Elements

We also analysed the frequency distribution of BPMN elements concerning the total amount of elements present in the entire collection of models. In particular, the 25,590 models are composed by 1,038,084 BPMN elements. The frequency

[11] As it can be seen from Fig. 1 we associated an ID to each BPMN element so, for presentation purposes, the following analysis and diagrams use the IDs instead of the plain name of the elements.

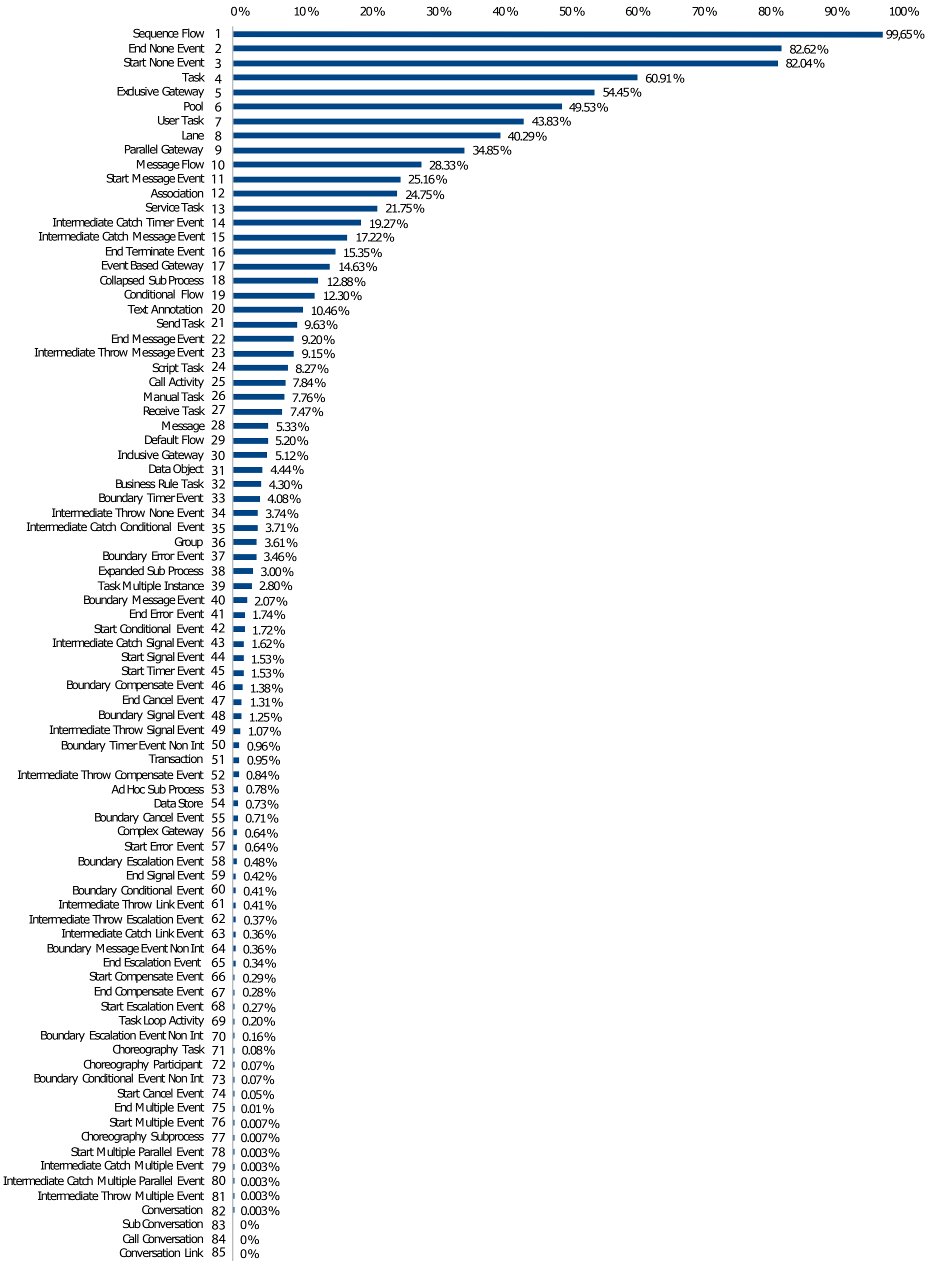

Fig. 1: Distribution of BPMN elements over models

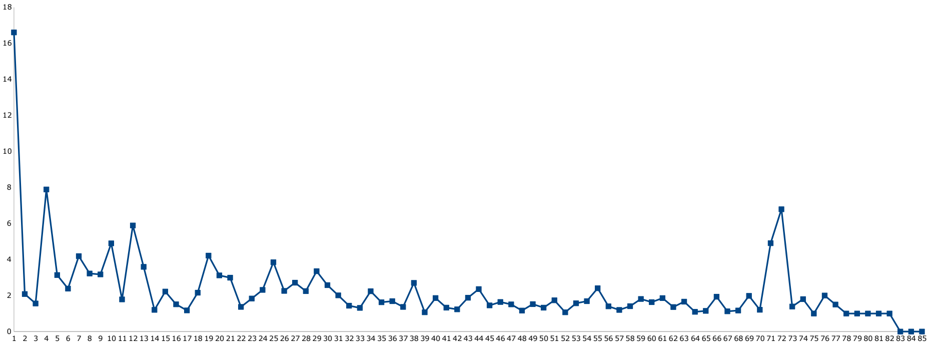

Fig. 2: Average number of occurrence of a given construct in a model

distribution of all BPMN elements is reported in Fig. 3, and it is ordered by following the ranking emerged from Fig. 1. Starting from the left to the right of the figure, we find rank 1, which corresponds to the number of Sequence Flow, rank 2, which corresponds to End None Event, till rank 85, which corresponds to Conversation Link. Besides the frequency distribution of BPMN elements, we also reported the Zipfian distribution (highlighted in red). The Zipfian distribution states that the frequency of words in natural languages is inverse to their rank [24]; this juxtaposition allows us to highlight that BPMN exhibits a very close distribution to that one of word usage in natural languages.

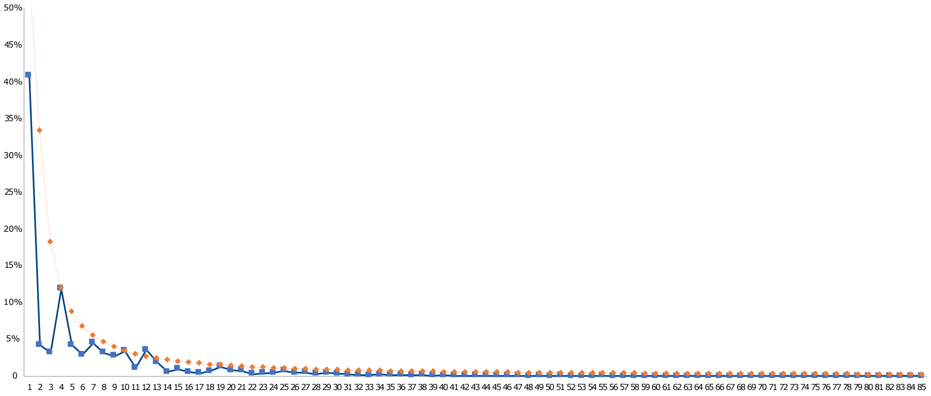

Fig. 3: Frequency plot of BPMN elements by rank

The 85 BPMN elements, according to the standard, can be grouped in eight families: *Activities*, *Gateways*, *Events*, *Connecting Objects*, *Swimlanes*, *Artefacts*, *Choreography* and *Conversation*. Therefore, to provide additional insight on the frequency distribution of BPMN elements, we reported in Fig. 4 statistics over their usage grouped by families; we also distinguished between categories in which the families can be organised. Where possible, we distinguished between basic elements such as Normal Task, Start/End None events, and their *Typed* versions, which allow a modeller to represent additional information like the fact that a task can be carried out manually or in an automatic way. To group additional elements of a family that are not part of a specific category, we used the term *Others*.

In Fig 4, we can notice that the most prominent family is the one of *Connecting Objects* that represents almost half of the used elements (49.51%), of which mostly are Sequence Flow. Concerning the *Activities* family (22.34%), most of the elements are Tasks (92.39%), more than half of which are Normal Tasks (57.37%); Sub-processes (3.96%) and other (3.65%) activities are less present. About the *Events* family, we notice a predominance of End (41.17%) and Start (35.16%) Events; most of which are of the basic form *None* (respectively 80.52% and 70.03%). The other events following are Intermediate (19.28%) and Boundary (4.38%) events. Intermediate events are divided into Throwing (28.85%) and Catching (71.05%), instead Boundary events are divided into Interrupting (91.26%) and Non-Interrupting (8.74%). Regarding the *Gateways* family, the Exclusive Gateway is the most used (54.63%), followed by the Parallel Gateway (35.37%) and the Others (10%). Elements of the *Swimlanes* family forms the 6.11% of the total and are divided into Pool (47.62%) and the Lanes (52.38%). The *Artefacts* family forms the 1.5% and most of its elements are Text Annotations (53.62%).

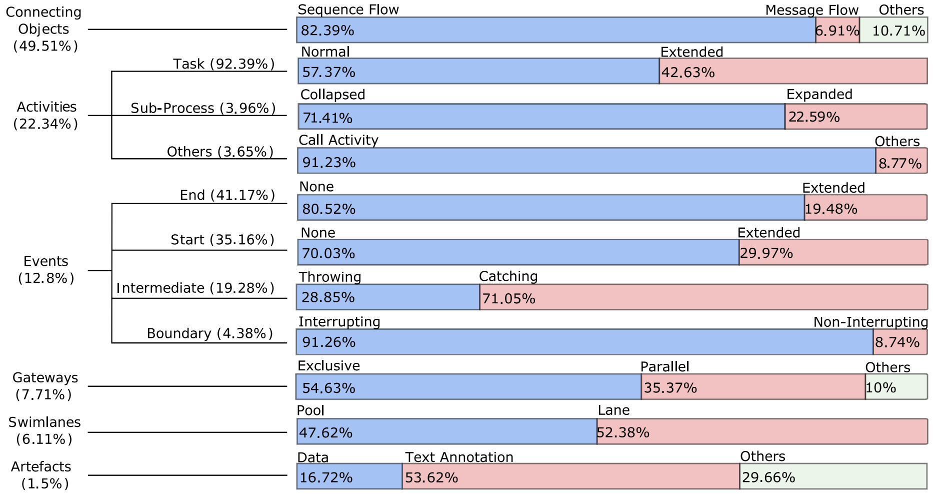

Fig. 4: Use of BPMN elements grouped by categories

4 Complexity of BPMN Models

In this section, we focus on the amount and the types of elements in a model.

4.1 BPMN Models Size

It is well known that the size of a BPMN model (i.e. the number of elements that form the model) affects its understandability [17]. Regarding the 25,590 models that we analysed, we detected that a model presents between 40-41 elements on average, with a median of 31 elements, a standard deviation of 63-64 elements, a maximum number of 3,672 elements and a minimum number of 8 elements. To provide an overview, we classified models based on their size, and we reported those data in Table 2. We divided models into three macro-sets: from 8 to 100, from 101 to 200 and from 201 to over 2000. As we can notice, the first set presents the 96.22% models while the second and third sets respectively present 2.76% and 1.02% of models. The majority of the models (68.61%) are distributed in the classes from 11-20 to 41-50. In particular, the higher concentration is in class 11-20, which presents the 25.5% of models.

8-100			101-200			201-2000+		
Classes	N°of Models	% of Models	Classes	N°of Models	% of Models	Classes	N°of Models	% of Models
8-10	2 293	8.96	101-110	117	0.46	201-300	149	0.58
11-20	6 615	25.85	111-120	116	0.45	301-400	64	0.25
21-30	3 793	14.82	121-130	115	0.45	401-500	18	0.07
31-40	3 606	14.09	131-140	61	0.24	501-600	8	0.03
41-50	3 543	13.85	141-150	89	0.35	601-700	5	0.02
51-60	1 259	4.92	151-160	52	0.20	701-800	2	0.01
61-70	1 827	7.14	161-170	49	0.19	801-900	3	0.01
71-80	1 128	4.41	171-180	39	0.15	901-1001	2	0.01
81-90	355	1.39	181-190	43	0.17	1001-2000	4	0.01
91-100	202	0.79	191-200	25	0.1	2000+	8	0.03

Table 2: Number and percentage of models by size

4.2 Syntactic Complexity of BPMN Models

Studies conducted over other graphical notations (i.e., UML [20, 15]) reported that the theoretical complexity of the language (measured by the total number of the modelling elements) is different from the practical complexity given by the number of elements used in the model. To measure the practical complexity of the BPMN modelling notation, we extracted for each model the number of different notation elements used. In Fig. 5 we report the syntactic complexity of BPMN models. As we can see, even considering that the BPMN meta-model describes 85 elements, in our collection of models, the majority (over 54%) uses between 5 and 8 types of elements. Less than 1% of the models use more than 18 different types of BPMN elements. The calculated average of elements types present in a model is 8.85, meaning that a model is designed by using around 8 or 9 types of elements.

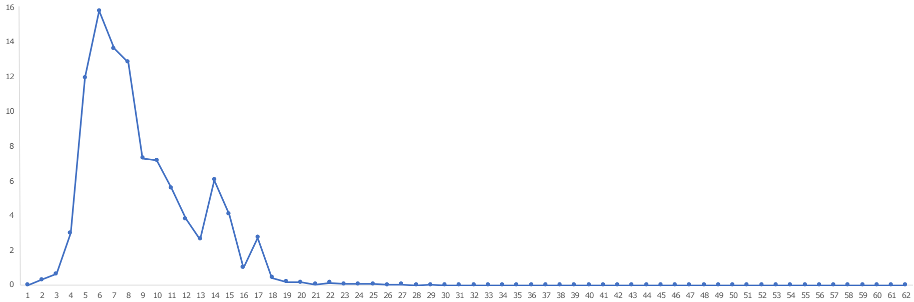

Fig. 5: Syntactic complexity of BPMN models

4.3 Variety of BPMN Subsets

For inspecting the variety of BPMN elements used in models, we used the Hamming Distance. The Hamming Distance refers to the number of different characters in two strings [13]. To calculate the Hamming Distance between two models, we mapped them into binary strings where positive bits signal the presence of a specific BPMN element while negative bits represent their absence. Therefore, for each model, we defined an 85-bit binary string that indicates (using 0s and 1s) which elements each model presents. Then we calculated the Hamming Distance between pair of strings (i.e. between pair of models), and we reported our findings in Fig. 6. As we can notice from the figure, a small percentage (0.66%) of models present a Hamming Distance of zero, which means the same type of BPMN elements forms the models; this highlights a high variability in usage BPMN notation. Most of the models (around 60%) differ from each other for 6-12 BPMN elements. While really few models (0.57%) differ for more than 20 elements, this is also since few models present such an amount of elements. The calculated average dissimilarity between two BPMN models is 8.5, meaning that a model's vocabulary differs from another by 8 or 9 types of elements.

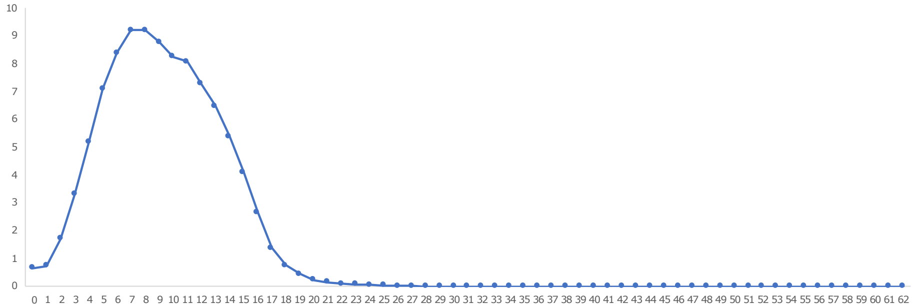

Fig. 6: Hamming distance of BPMN vocabularies

5 BPMN Elements Correlations

To check whether specific BPMN elements are used in combination, we analysed the possible correlation between pairs and groups of elements.

5.1 Correlation Between Pairs of BPMN Elements

To analyse pairs of elements, we determined the covariance matrix of all modelling elements - a square matrix giving the covariance between each pair of elements. Then we adopted the Pearson correlation coefficient (ρ), which corresponds to the covariance between two modelling elements divided by the product of their standard deviations, to normalise the covariance measurement to a value ranging between -1 and 1. Negative values represent inverse correlations, while positive values represent direct correlations. When the value is zero, no correlation is present; while the more the value is close to ±1, the stronger the correlation. In our analysis we distinguished between three degree of correlation: *small* when the value lies below ±0.29, *medium* when the value lies between ±0.30 and ±0.49, *strong* when the value lies between ±0.50 and ±1. For presentation purposes, in Table 3 we report only the most significant pairs of elements that presented a strong direct correlation. None of the elements pairs presented a medium/strong inverse correlation. The pairs presenting a small inverse correlation indicate typed elements at the expense of the not typed ones. It is the case of a model presenting typed elements such as Send Task, Receive Task, Manual Task, Business Task may not present the not typed Task element. Moreover, rarely used elements presented a small inverse correlation with many of the other elements.

For what concerns the *Strong correlations*, we report in the following our interpretation of the obtained results. The highest correlation value corresponds to $\rho = 0.87$, and it is obtained by the pair *Exclusive Gateway-Sequence Flow*; Exclusive Gateways are used to split the workflow into multiple branches using multiple Sequence Flows. It is reasonable, therefore, to notice this strong correlation. *Send Task* and *Receive Task* present a correlation coefficient of $\rho = 0.79$. Those elements are used to model communication aspects regarding a single or multiple communicating processes, and their strong correlation is straightforward. The pair composed by *Start Event None* and *End Event None* have a cor-

Element One	**Element Two**	ρ
Exclusive Gateway	Sequence Flow	0.87
Send Task	Receive Task	0.79
Start Event None	End Event None	0.78
Sequence Flow	Task	0.71
Expanded SubProcess	Start None Event	0.67
Inclusive Gateway	Exclusive Gateway	0.65
Exclusive Gateway	Task	0.57
Expanded SubProcess	End None Event	0.56
Pool	Message Flow	0.53
Pool	Lane	0.52

Table 3: Correlation coefficient between BPMN elements

relation coefficient of $\rho = 0.78$. Generally, a business process model has at least one start event, and one end event explicitly reported and the not typed once (those named *None*) are the most used. The pair composed by *Sequence Flow* and *Task* presents a strong correlation ($\rho = 0.71$) since tasks are key elements in the design of a business process model, and sequence flows are attached to tasks for defining the control flow. Being and *Expanded Sub-Process* a business process per se, it is also reasonable that its start and end are represented and strong correlation with *Start None Event* and *End None Event*. It results a ρ value of 0.67 and 0.56, respectively. The *Inclusive Gateway* and the *Exclusive Gateway* elements present a correlation coefficient $\rho = 0.65$; this means that the Inclusive Gateway, when used, it is used in combination with other gateways and being the Exclusive Gateway the most used one, the correlation is straight forward. The pair *Exclusive Gateway-Task* presents a correlation coefficient $\rho = 0.57$; this is expected since both elements are among the most used ones. The *Pool* element that is generally used to represent the Owner of a process (e.g. a Company) is often (but not always) used in combination with *Lanes* to distinguish between departments of the same organisation. Therefore a strong correlation with the element Lane ($\rho = 0.52$) has to be expected. The *Pool* element is also used for representing collaborations of processes which generally corresponds to different companies collaborating to achieve some goal. Therefore, it is generally used in combination with the *Message Flow* ($\rho = 0.53$) to represent the exchange of messages between the different parties involved in the process.

5.2 The Combined Use of BPMN Elements

After analysing the correlation coefficient for pairs of BPMN elements, we analysed groups of elements to find those most frequently used in combination.

The Venn diagram in Fig. 7 shows the elements that are most used in combination with the respective percentage of the models in which they are present. We reported only those combinations of elements that appeared together in at least 25% of the models. The dashed lines size are used to group the elements so that the shorter the dash, the higher the percentage of models with that combination of elements. The elements that are primarily used in combination are Tasks and Sequence Flows with a percentage of 96%. Instead, Start Event and the End Event are present in combination in 91% of the models. The above mentioned elements (i.e., Task, Sequence Flow, Start and End Event) form a core set of elements that are mostly used in combination; they are present together in 89% of the models. Additional elements combined with the core set are delimited with a green dashed line. The combination of the core set with Exclusive Gateway is present in 50% of the models, with Pool in 44% of the models, with Parallel Gateways present in 31% of the models, with Intermediate Events is present in 30% of the models. On the right side of the figure, we reported the combination of Lanes and Message Flows with the Pool element, which respectively are present in combination in 37% and 27% of the models. In particular, these elements are related to the modelling of Organisational aspects (i.e., the owner of the process represented by a Pool of other company departments by

Lanes) and the modelling of processes collaborations (i.e., Pools with processes that communicate via message flows). It is worth noticing that typed elements (e.g, User Task, Message Event, Timer Event, etc.) are not included in any of the most popular BPMN element subsets.

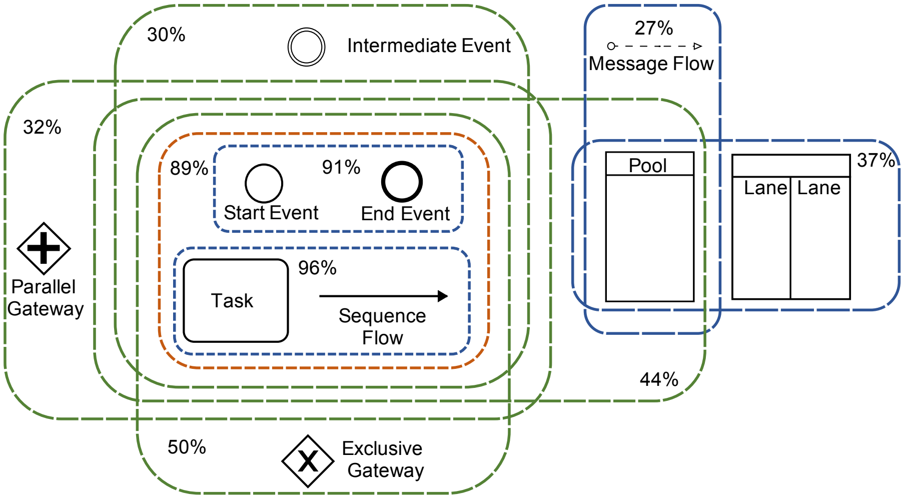

Fig. 7: Most popular BPMN vocabulary subsets

6 Comparison with Muehlen and Recker's work

Among the literature targeting BPMN, the contribution that most relates to ours is [18]. The authors analysed the usage of BPMN v1.0 over a set of 120 models. The BPMN version that is currently used is BPMN v2.0 which includes new elements with respect to version one. In fact, in our study, we extract data for 85 BPMN elements respect to the 50 elements described in their work. Also, the significant amount of models that we retrieved, 25,590, compared to their 120 models, establishes a stronger base for conducting statistical analysis.

For what concerns the distribution of the occurrence of elements, both the research works highlight a similar trend in the usage of BPMN elements concerning Sequence Flows, Start and End None Events, Tasks, Exclusive Gateway, Pool, Lane, Parallel Gateway and Message Flow being among the most used elements of the notation. The differences in the rankings are evident when we look at those elements that are less used (Multiple Instance, Cancel Events, etc.). However, being among the less used elements, the ranking highly depends on the set of analysed models.

Regarding the frequency distribution of BPMN elements divided by ranks, both the analysis of [18] present a distribution that resembles the power law distribution typical of natural languages (Zipf's law [24]). Thus, we observe that despite the elements added in v2.0 the BPMN language still maintains the frequency distribution of typical natural languages, consisting of a few essential elements and a set of barely used ones.

The set of models we analysed presents a higher variety of elements used, with a maximum of 62 and a higher concentration of models that are composed by a number between 5 and 8 different elements; in [18] the maximum is 15 different elements with the majority of models composed by a number between 6 and 12 different elements. In any case, we confirmed that on average a model is designed with a number between 8 and 9 different types of BPMN elements. In [18], looking at the results obtained from the calculation of the Hamming Distance of BPMN vocabularies, the authors detected a maximum difference of 18 types of elements with the majority of the models differing for 7-8 types of elements. In addition, the authors stated that this value might decrease in the future as wider adoption of BPMN may result in a more homogeneous use of BPMN vocabulary for designing models. Instead, in our results, we obtained a maximum hamming distance of 62, with the majority of the models differing for 6-12 types of elements, therefore assisting to an increase in the variety of BPMN elements used.

Referring to correlations between elements of the notation, our findings do not show any significant (medium and strong) inverse correlation that could lead us to conclude that some elements are actually used alternatively. At the same time, in [18] some inverse correlations are highlighted, but they report a weak interpretation. Instead, for what concerns medium and strong direct correlations, we identified many more correlations to them. However, some correlations they detected did not result from our analysis (e.g., Lane and Message Flow, Start Message and Exclusive Gateway, Start Message and End Terminate).

Referring to the combined usage of BPMN subsets, we obtained results with the same trend of [18]. The pair Task-Sequence Flow it constitutes, for both, the most used combination present in 96% of the analysed models. For all the other combinations we obtained higher percentages denoting a more tight usage of those subsets. Those differences may be due to the different amount of analysed models (i.e., we analysed 25,590 models while they analysed only 120). In addition, compared to [18], we identified more subsets that are used in combination, especially the subsets formed by Intermediate Events and the core set and the subset formed by Message Flows and Pool, that did not emerge from the previous contribution. Finally, in Table 4 we report the comparison of the analyses with Muehlen and Recker's work described above.

	Our work	**[18]**
Number of BPMN models repositories	6	3
Number of analysed BPMN models	25 590	126
Average complexity of BPMN models	8-9 elements	6-12 elements
Average variety of elements in models	6-12 elements	7-8 elements
Average number of construct's occurrence	✓	✗
BPMN elements distribution over models	✓	✓
BPMN elements frequency distribution	✓	✓
BPMN elements correlations	✓	✓
Combined use of BPMN elements	✓	✓

Table 4: Comparison of this work with Muehlen's work [18]

[1] Janiesch, C., Koschmider, A., Mecella, M., et al. (2017). The internet-of-things meets business process management: mutual benefits and challenges, arXiv:1709.03628, 2017.

7 Conclusion, Limitations and Future Work

This research work's objective was to investigate whether the BPMN vocabulary adopted nowadays by model designers shows some particular trends. The main findings confirm that the majority of the models designed with BPMN 2.0 are designed around a core set of elements (i.e., Task, Sequence Flow, Start Event and End Event), resulting in a large portion of the notation that is rarely used. These findings are consistent with those obtained for BPMN 1.0 [18]. These findings are confirmed by comparing the BPMN theoretical complexity, which corresponds to 85, and the practical complexity of the notation, which corresponds to 8-9 elements on average. Our study clearly emphasises the wide usage of the core set of BPMN elements with respect to the most advanced ones. In addition, our results show that some elements are highly correlated, and some are often used in combination; this also can be taken as a reference while preparing training sessions on BPMN. The emerged results can be taken as a reference for guiding the development of BPMN-related tools, which should focus first on providing support for the most used elements and then for the rest of the notation. The results can also affect training activities suggesting a list of elements (the most used ones) that trainers should focus on first before addressing advanced elements (the less used ones).

It should be stated that this research was subject to a limitation given by the analysed models. However, we argue that 25,590 models taken from six different repositories should compose a solid base for generalising the obtained results to general use of BPMN notation.

In the future, we plan to extend our study to incorporate additional models possibly coming from real world applications or additional online repositories. We also want to analyse the usage of the BPMN notation by targeting different application domains to discover possible subsets of BPMN elements that better fit an application domain with respect to another. Finally, we envision a study that involves practitioners and fresh BPMN users to evaluate whether and how the experience, gained by practising BPMN, may lead to a differentiated usage of BPMN elements or whether it is the application domain mostly guides the choice of the BPMN elements to use.

References

1. Aguilar, E.R., Cardoso, J.S., García, F., Ruiz, F., Piattini, M.: Analysis and validation of control-flow complexity measures with BPMN process models. In: BPM, Development and Support. LNBIP, vol. 29, pp. 58–70. Springer (2009)
2. Bork, D., Karagiannis, D., Pittl, B.: Systematic analysis and evaluation of visual conceptual modeling language notations. In: Research Challenges in Information Science. pp. 1–11. IEEE (2018)
3. Bork, D., Karagiannis, D., Pittl, B.: A survey of modeling language specification techniques. Information Systems **87** (2020)
4. Compagnucci, I., Corradini, F., Fornari, F., Polini, A., Re, B., Tiezzi, F.: Modelling Notations for IoT-Aware Business Processes: A Systematic Literature Review. In: BPM Workshops. LNBIP, vol. 397, pp. 108–121. Springer (2020)
5. Corradini, F., Ferrari, A., Fornari, F., Gnesi, S., Polini, A., Re, B., Spagnolo, G.O.: A guidelines framework for understandable BPMN models. DKE **113**, 129–154 (2018)
6. Corradini, F., Fornari, F., Polini, A., Re, B., Tiezzi, F.: A formal approach to modeling and verification of business process collaborations. SCP **166**, 35–70 (2018)
7. Corradini, F., Fornari, F., Polini, A., Re, B., Tiezzi, F.: Reprository: a repository platform for sharing business process models. CEUR, vol. 2420, pp. 149–153 (2019)
8. Dijkman, R.M., Dumas, M., Ouyang, C.: Semantics and analysis of business process models in BPMN. Information of Software Technology **50**(12), 1281–1294 (2008)
9. Fahland, D., Favre, C., Jobstmann, B., Koehler, J., Lohmann, N., Völzer, H., Wolf, K.: Instantaneous soundness checking of industrial business process models. In: Business Process Management. LNCS, vol. 5701, pp. 278–293. Springer (2009)
10. Genon, N., Heymans, P., Amyot, D.: Analysing the cognitive effectiveness of the bpmn 2.0 visual notation. In: Software Language Engineering. LNCS, vol. 6563, pp. 377–396. Springer (2010)
11. Haisjackl, C., Pinggera, J., Soffer, P., Zugal, S., Lim, S.Y., Weber, B.: Identifying quality issues in BPMN models: an exploratory study. In: Enterprise, Business-Process and Information Systems Modeling. LNBIP, vol. 214, pp. 217–230. Springer (2015)

12. Haisjackl, C., Soffer, P., Lim, S.Y., Weber, B.: How do humans inspect BPMN models: an exploratory study. Software of System Modeling **17**(2), 655–673 (2018)
13. Hamming, R.: Error detecting and error correcting codes. vol. 29, pp. 147–160. Nokia Bell Labs (1950)
14. Heinze, T.S., Stefanko, V., Amme, W.: Mining BPMN processes on github for tool validation and development. LNBIP, vol. 387, pp. 193–208. Springer (2020)
15. Kobryn, C.: UML 2001: A standardization odyssey. vol. 42, pp. 29–37 (1999)
16. Kossak, F., Illibauer, C., Geist, V., Kubovy, J., Natschläger, C., Ziebermayr, T., Kopetzky, T., Freudenthaler, B., Schewe, K.: A Rigorous Semantics for BPMN 2.0 Process Diagrams. Springer (2014)
17. Mendling, J., Sánchez-González, L., García, F., Rosa, M.L.: Thresholds for error probability measures of business process models. J. Syst. Softw. **85**(5), 1188–1197 (2012)
18. zur Muehlen, M., Recker, J.: How much language is enough? theoretical and practical use of the business process modeling notation. In: Advanced Information Systems Engineering, pp. 429–443. Springer (2013)
19. OMG: Business Process Model and Notation (BPMN), Version 2.0.2 (2013), https://www.omg.org/spec/BPMN/2.0.2
20. Siau, K., Erickson, J., Lee, L.: Theoretical vs. practical complexity: The case of UML. vol. 16, pp. 40–57 (2005)
21. Wohed, P., van der Aalst, W.M.P., Dumas, M., ter Hofstede, A.H.M., Russell, N.: On the suitability of BPMN for business process modelling. In: Business Process Management. LNCS, vol. 4102, pp. 161–176. Springer (2006)
22. Wong, P.Y.H., Gibbons, J.: Formalisations and applications of BPMN. Science of Computer Programming **76**(8), 633–650 (2011)
23. Zarour, K., Benmerzoug, D., Guermouche, N., Drira, K.: A systematic literature review on BPMN extensions. Business Process Management **26**(6), 1473–1503 (2020)
24. Zipf, G.K.: On the dynamic structure of concert-programs. vol. 41(1), pp. 25–36 (1946)