

# Process-To-Text: a framework for the quantitative description of processes in natural language

Yago Fontenla-Seco<sup>1</sup>[0000-0002-7566-0164], Manuel Lama<sup>1</sup>[0000-0001-7195-6155],  
and Alberto Bugarín<sup>1</sup>[0000-0003-3574-3843]

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)  
Universidade de Santiago de Compostela, Spain.  
{yago.fontenla.seco, manuel.lama, alberto.bugarin.diz,}@usc.es

**Abstract.** In this paper we present the Process-To-Text (P2T) framework for the automatic generation of textual descriptive explanations of processes. P2T integrates three AI paradigms: process mining for extracting temporal and structural information from a process, fuzzy linguistic protoforms for modelling uncertain terms, and natural language generation for building the explanations. A real use-case in the cardiology domain is presented, showing the potential of P2T for providing natural language explanations addressed to specialists.

**Keywords:** Process mining · Natural Language Generation · Explainable AI

## 1 Introduction

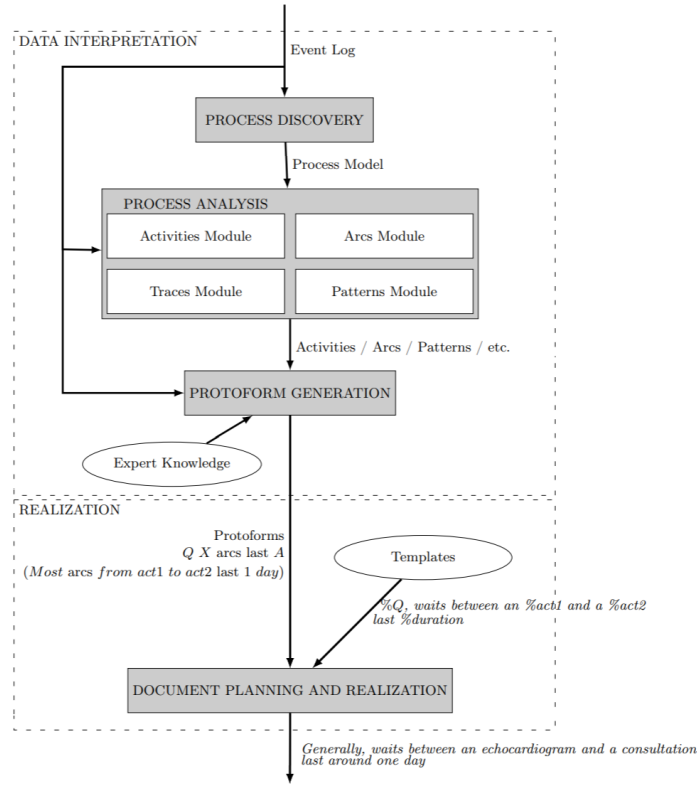
Processes constitute a useful way of representing and structuring the activities in information systems. The Process Mining field offers techniques to discover, monitor and enhance real processes extracted from the event logs generated by processes execution, allowing to understand what is really happening in a process, which may be different from what designers thought [2]. Process models are usually represented with different notations that represent in a graphical manner the activities that take place in a process as well as the dependencies among them. Process models tend to be enhanced with properties such as temporal information, process execution-related statistics, trends of process key indicators, interactions between the resources involved in the process execution, etc. Information about these properties is shown to users through visual analytic techniques that help to understand what is happening in the process execution from different perspectives. However, in real scenarios process models are highly complex, with many relations between activities, which make them nearly impossible to be interpreted and understood by the users [2]. Furthermore, the amount of information that can be added to enhance the process description is also very high, and it is quite difficult for users to establish its relation with the process model. Additional analytics which summarize quantitative relevant information about a process, such as frequent or infrequent patterns, that make it easier to focus on finding usual or unexpected behaviors, are also very useful. [4]

Nevertheless, their correct interpretation is usually difficult for users, since they need to have a deep knowledge about process modelling and process related visual analytics. In this regard, some approaches have been described to automatically generate natural language descriptions of a process aiming to provide users with a better understanding of it [10]. These descriptions aim to explain in a comprehensible way, adapted to the user information needs, the most relevant information of the process. In general, textual information is complementary to process model visualization, which is the usual way of conveying information to users.

Natural Language Generation (NLG) [6, 16] provides different methods for generating insights on data through natural language. Its aim is to provide users with textual descriptions that summarize the most relevant information of the data that is being described. Natural language is an effective way of conveying information to humans because i) it does not rely on human capability to identify or understand visual patterns or trends; and ii) it may include uncertain terms of expressions, which are very effective for communication [13]. Research suggests that in some specialized domains knowledge and expertise are required to understand graphical information [12] and proves that specialists can take better decisions based on textual summaries than on graphical displays [8]. With the integration of NLG and process model and analytics visualization, users with limited expertise on process modeling and analysis are provided with an AutoAI tool that allows them to understand the relevant information about what is really happening within a process.

In the state-of-the-art, techniques for process textual description focus on two perspectives. On the one hand, the Control-Flow perspective aims to align a Business Process Model Notation (BPMN) representation of a process model with its corresponding textual description [1, 10, 11, 17]. The aim of this proposal is to, first, facilitate users to understand the process model structure and the dependencies between activities and how resources are used and second, detect inconsistencies in the process model such as missing activities, activities in different order, or misalignments in order to maintain a stable process model through the different stakeholders of an organization. The main drawback of these approaches is that they focus on hand-made processes (not discovered from real data) with a well-defined structure, and draw all their attention in the validation step of the process design phase. Preventing its application to process models extracted from real-life data, usually unstructured, with many relations between activities, including frequent loops, parallels and choices. On the other hand, Case Description techniques focus on generating textual descriptions about the execution of single activities or activity sequences that have been registered in an event log [5]. The underlying process model is not discovered from the event log, therefore neither the process structure is considered nor the relations between activities. So, these last techniques do not provide a faithful description of what has happened during the process execution.

In this paper, we present the Process-To-Text (P2T) framework, a process mining-based framework (the real process model is discovered from the event



**Fig. 1.** Framework for the linguistic description of processes

data) for the automatic generation of natural language descriptions of processes. Descriptions include information from both the control-flow, case and specially time perspectives, the later being usually neglected in the literature. P2T is based on a Data-to-Text (D2T) architecture [15] using linguistic protoforms (as a way to handle imprecision) that will be generated into natural language texts following a hybrid template-based realization approach.

## 2 P2T: a framework for textual description of processes

Fig. 1 depicts the Process-To-Text (P2T) framework, for the automatic generation of textual descriptive explanations of processes. This framework is based on the most widely used architecture for D2T systems [15], which defines a pipeline composed of four stages: *signal analysis*, *data interpretation*, *document planning*, and *microplanning and realization*. P2T does not include the *signal analysis* stage since data input are not numerical, but event logs. Also, the *document planning* stage is subsumed in the *microplanning and realization* stage.

**Process discovery.** The input of the data interpretation stage is an *event log*, defined as a multiset of traces. A trace is a particular execution of a process (i.e., a case), and it is represented as an ordered sequence of events, being an event the execution of an activity, which contains context information about the execution said activity (e.g. timestamp, the case it belongs to, the resources involved in its execution, the process variables modified, etc.). However, the event log itself is not used as a direct input to generate the descriptions. Firstly, the process model has to be discovered from the event log by applying a process mining algorithm [2, 3], which, traditionally have followed heuristic [19], inductive [9], or evolutionary computation [18] based approaches.

**Process Analysis.** Once the process is mined, the log is replayed (each trace is played over the discovered model) [2]. This gives us both temporal and frequency-based information about activities, arcs (relations between activities) and traces that can be as well used to extract frequent and infrequent behavioral patterns [4]. Then, this information is summarized into indicators (e.g. average duration and frequency of the relation between activities, average and mode duration of a path, changes of mean duration of an activity within a period, etc.) which are computed in the modules depicted in Figure 1 in the process analysis phase. This phase is part of the framework and indicators are computed for any case or domain.

**Protoform generation.** A protoform [20] is an abstracted summary which serves to identify the semantic structure of the object to which applies. In P2T, protoforms include fuzzy temporal references for providing information about the temporal dimension of activities, arcs, or traces. For example, the textual description *Most executions of activity  $\alpha_1$  last ten minutes in average more than those of activity  $\alpha_2$*  is generated by the protoform *Q B activity lasts A*. This is an activity-related protoform where *Q* is the quantifier *Most* *B* is a qualifier, in this case, activity  $\alpha_1$ , and *A* is the summarizer used to describe activity  $\alpha_1$  *ten minutes in average more than those of activity  $\alpha_2$* . Note that behavioral patterns can be expressed through relations between activities, meaning that pattern-related protoforms are compositions of arcs-related protoforms. Protoforms have abstraction levels, as summarization does, this allows for a general abstracted summary to produce multiple different summaries depending on the knowledge used for its realization.

**Document planning and Realization** Its objective is to generate the natural language descriptions of the process, taking the protoforms, templates, and expert knowledge as inputs. In our model we follow a hybrid template-based realization, which uses some domain expert knowledge and is more rich and flexible than basic fill-in-the-gap approaches, but simpler and quicker than full fledged NLG system implementations. The SimpleNLG-ES [14] (Spanish version of the SimpleNLG realization engine) realization engine is used in this stage.

### 3 Case study

We have applied our framework in a real case study in the health domain: the process related to the patients' management in the Valvulopathy Unit of the Cardiology Department of the University Hospital of Santiago de Compostela. In this Unit, consultations and medical examinations, such as radiography, echocardiogram or Computed Tomography (CT) scans, are performed to patients with aortic stenosis in order to decide the treatment (including surgery) they will undergo. Other information like unexpected events (e.g. emergencies, non-programmed admissions) and patient management activities (inclusion in the process, revisions, etc) are also recorded in the event log data.

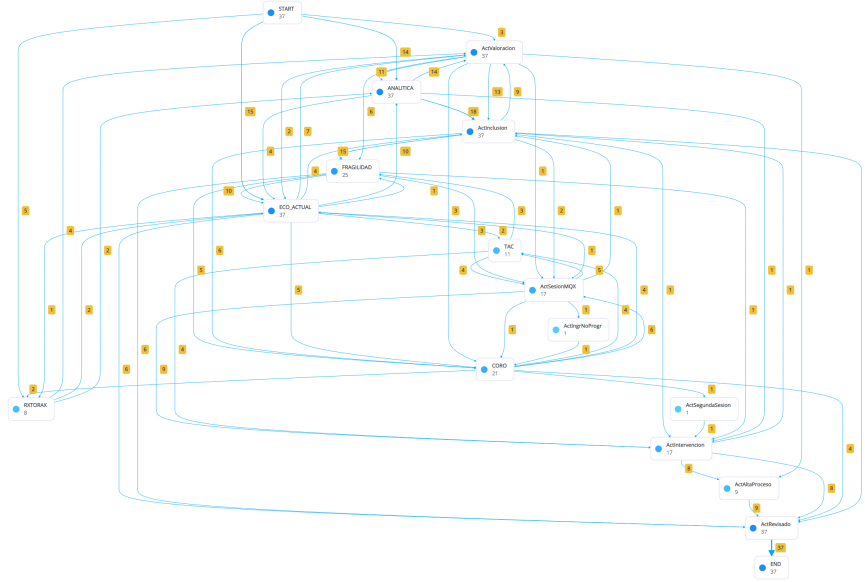
Medical professionals have a real interest in applying process mining techniques to this process, since it allows them to extract valuable knowledge about the Unit like, relations between age, sex, admittance (emergency or normal admission) and the number of successful surgeries or delay between activities, the delays between crucial activities (such as the admission of a patient and its surgery) due to tests like CT scans or echocardiograms, the different paths of the process which patients with different attributes follow, etc. The main goal to reach with all this information is to reduce the delays between process activities, prevent the repetition of activities (loops in the process), minimize patient management time, optimize resources and most important, increase the number of successful treatments.

In Fig. 2 the model that describes this process is shown, it has been discovered from an event log with data on 639 patients.

Since medical, management activities and exceptions are recorded, and since patients' management depends on their pathological state, the frequency of each path (sequence of activities on a trace) in the process is very low (only the twenty most frequent paths from the six-hundred twenty-two total are shown in the figure) giving place to a highly complex model. This makes it difficult for medical professionals to understand what happens within the process even when it is graphically represented. To solve this, linguistic descriptions of the main process analytics are generated, as shown in Table 1, facilitating the understanding of temporal relations and delays between activities and traces, which is the main concern of domain experts in this regard.

### 4 Conclusions and Future Work

Our P2T framework integrates the process mining, natural language generation, and fuzzy linguistic protoforms paradigms for the automatic generation of textual descriptive explanations of processes, which include quantitative information (i.e., frequent and infrequent behavior, temporal distances between events and frequency of the relationships between events). A real use-case is presented, showing the potential of P2T for providing natural language explanations addressed to cardiology specialists about consultations and interventions of the patients of the valvulopathy unit. As future work, extensive human validation of the generated descriptions will be conducted with domain experts.



**Fig. 2.** Model of the Valvulopathy process represented with the InVerbis Analytics visualization tools [7].

**Table 1.** Textual descriptions generated for the valvulopathy process

- 
- During the first half of year 2018, 52% less Surgical Interventions were registered compared to the second half of that same year. period.
  - In the process, 78% less Surgical Interventions than Coronographies were registered.
  - Waiting time between Consultations is around 5 weeks and 6 days in average.
  - Waiting time between a Coronography and a CAT is around 6 weeks in average.
  - Around 7 weeks and 3 days after the Medical-Surgical Session a patient undergoes Surgical Intervention.
  - 6% of times, after the First Medical-Surgical Session, a Second Session is held around 5 weeks and 3 days later. On the contrary, 33% of times, patient undergoes Surgical Intervention around 7 weeks and 3 days later.
  - Patients who go through Assessment, Medical-Surgical Session and Surgical Intervention stay for 114 days in average at the cardiology service.
- 

## Acknowledgments

This research was funded by the Spanish Ministry for Science, Innovation and Universities, the Galician Ministry of Education, University and Professional Training and the ERDF/FEDER program (grants TIN2017-84796-C2-1-R, ED431C2018/29 and ED431G2019/04).

## References

1. van der Aa, H., Leopold, H., Reijers, H.: Detecting inconsistencies between process models and textual descriptions. In: Proceedings 13th Int. Conf. on Business Process Management, BPM 2015. vol. 9253, pp. 90–105. Springer (2015)
2. van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer (2016)
3. Augusto, A., Conforti, R., Dumas, M., Rosa, M.L., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering* **31**(4), 686–705 (2019)
4. Chapela-Campa, D., Mucientes, M., Lama, M.: Mining frequent patterns in process models. *Information Sciences* **472**, 235–257 (2019)
5. Dijkman, R.M., Wilbik, A.: Linguistic summarization of event logs: A practical approach. *Information Systems* **67**, 114–125 (2017)
6. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.* **61**(1), 65–170 (Jan 2018)
7. InVerbis Analytics: <https://processmining.inverbisanalytics.com/>, [Online; accessed in February 2021]
8. Law, A.S., Freer, Y., Hunter, J., Logie, R.H., McIntosh, N., Quinn, J.: A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing* **19**(3), 183–194 (2005)
9. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.: Discovering block-structured process models from event logs - a constructive approach. In: Application and Theory of Petri Nets and Concurrency. pp. 311–329. Springer (2013)
10. Leopold, H., Mendling, J., Polyvyanyy, A.: Generating natural language texts from business process models. *Lecture Notes in Computer Science* **7328**, 64–79 (2012)
11. Leopold, H., Mendling, J., Polyvyanyy, A.: Supporting process model validation through natural language generation. *IEEE Trans. Soft. Eng.* **40**(8), 818–840 (2014)
12. Petre, M.: Why looking isn't always seeing: Readership skills and graphical programming. *Commun. ACM* **38**, 33–44 (06 1995)
13. Ramos-Soto, A., Bugarín, A., Barro, S.: Fuzzy sets across the natural language generation pipeline. *Progress in Artificial Intelligence* **5**(4) (2016)
14. Ramos-Soto, A., Janeiro-Gallardo, J., Bugarín, A.: Adapting SimpleNLG to spanish. In: 10th International Conference on Natural Language Generation. pp. 142–146. Association for Computational Linguistics (2017)
15. Reiter, E.: An architecture for Data-to-Text systems. In: 9th European Workshop on NLG (2007). pp. 97–104 (2007)
16. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, USA (2000)
17. Sánchez-Ferreres, J., Carmona, J., Padró, L.: Aligning textual and graphical descriptions of processes through ilp techniques. In: Advanced Information Systems Engineering. pp. 413–427. Springer International Publishing, Cham (2017)
18. Vázquez-Barreiros, B., Mucientes, M., Lama, M.: Prodigen: Mining complete, precise and minimal structure process models with a genetic algorithm. *Information Sciences* **294**, 315–333 (2015)
19. Weijters, A., Aalst, W.: Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering* **10**, 151–162 (07 2003)
20. Zadeh, L.A.: A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. In: Proc. Annual Meeting North American Fuzzy Information Processing Society (NAFIPS-FLINT 2002). pp. 523–525 (2002)