Software Size Estimation in the Early Stages of Development

M. Elena Centeno[1], Mabel Bertolami[2], Alejandro Oliveros[3]

[1]Departamento de Informática, Facultad de Ingeniería, UNPSJB, Argentina
malenac163@gmail.com
[2]Departamento de Informática, Facultad de Ingeniería, UNPSJB, Argentina
mbertolami@gmx.net
[3]Magíster de Ingeniería de Software, Facultad de Informática, UNLP, Argentina
aoliveros@gmail.com

Abstract. This paper presents a pilot study that raised the estimate of functional size in the very early stages of software development. The goal is to find a relationship between some measures of artifacts produced in the Requirements Elicitation process such as the Language Extended Lexicon (LEL) and Function Points (FP) of the scenarios. From a set of case studies the respective LELs were measured and the FP of the scenarios were calculated. The values obtained were related using different regression models. The models were object of a rigorous statistical analysis to identify the one that best represents the experimental data based on the values of their statistics. Finally, the model was tested on data collected from three applications that are implemented.

Keywords: Language Extended Lexicon, Function Points, regression

1 Introduction

The size of the software plays an essential role for the estimation of the factors associated with a software development project (effort, staffing, schedule, cost, etc.) [11]. Function Point Analysis (FPA) is a technique widely used that can be applied from the requirements phase and throughout the entire lifecycle. To anticipate the measurement to stages previous to the definition of the requirements the procedure SFP (Scenario Function Points) [1] could be apply to estimate the functional size of the scenarios generated from the Language Extended Lexicon (LEL) [9].

The main contribution of this proposal is to estimate the functional size of the system to be developed in stages prior to the requirements. Because the scenarios are derived from the LEL using specific heuristics, one wonders, can further anticipate the estimation of Function Points (FP)? Is it possible to estimate the FP of a software system based on the size of the LEL?

The rest of this article is organized as follows: section 2 is a review of papers that propose methods for the calculation of FP in the early stages. In section 3 presents the basic concepts of the LEL and scenarios. In section 4 describes the measurements applied to LEL. In Section 5 summarizes the procedure for measuring the functional size. Section 6 includes data from the measurements. In Section 7 presents the statistical analysis that set a model for the estimation of FP from a measure of the LEL and Section 8 discusses the conclusions and future work.

2 Related Works

There are several proposals for the estimation of functional size in the earliest phases of software development. Some can be applied from the stage of feasibility study (for example, based on the data repository ISBSG [13], Early & Quick FP [14]), however, with the exception of the approach of Choi *et al*. [15], there were no other proposals for the products of the Requirements Elicitation.

Regarding the artifact used for the estimation, some proposals use the data model, a significant number uses the Requirements Specification and in recent times have been developed multiple approaches based on Use Cases, such as Use Case Points [16], the approach of Fetcke *et al*. [17].

In fact these approaches require documents with a high level of detail. This exceeds the amount of information available in the earliest stages of a project and in some cases not very different from what is required for measurement with a standard method. It is noteworthy that in the literature found no similar work to that proposed in this article.

3 LEL and Scenarios

The construction of scenarios is based on the vocabulary of the Universe of Discourse (UD). The LEL is a representation of the symbols in the language of the application and seeks to understand the vocabulary used by the user in his real world without worrying about understanding the problem. LEL entries are classified as Subject, Object, Verb and State [9].

The scenarios are used to understand the application and its functionality: each scenario describes a specific situation focusing on their behavior. The approach of scenarios that we follow [9] uses a template of the scenario compound by name, goal, context, actors, resources and episodes. Its construction is based solely on the information contained in the LEL.

The LEL and Scenarios (L&S) is a tool used in the stage of Requirements Elicitation. It permits to capture and derive requirements but it is not a requirements specification. In [5] and [9] one can find the references to the works on the subject.

4 Measures of Size of LEL

Among the measures that characterize the LEL, in this first proposal, some were chosen because they are direct measures so they are easy to obtain. First, the total number of symbols is used (proposed in [2]). Also the group *noun symbols*, formed by the symbols in the categories Subject and Object, the group *verb symbols* and the group *impacts of verb symbols* were added. The *impacts of verb symbols* were included because in the process of building scenarios, the impacts of a verb symbol included in the impact of a primary or secondary actor is used to describe episodes of the scenario [9]. As the episodes describe the behavior it was considered appropriate to study its relationship with the FP.

5 Measurement of Functional Size of Scenarios

The procedure SFP [1] allows estimating the size in FP of a software system from the scenarios. Its design is based on the method IFPUG FPA [6] and the structure proposed in the ISO / IEC 14143-1 [7] for functional size measurement methods.

The definition of SFP establishes conceptual mapping between the components of the scenarios and those proposed by IFPUG. The SFP model consists of three Functional Components: resources, External Input (EI) and External Output (EO). SFP includes rules for the identification and classification of types of entities, numerical assignment rules for valuing the entities of the model and a function to derive the functional size from individual components. The complexity and contribution in FP of EIs, EOs and resources is determined in a similar way using the IFPUG tables for EI, EQ and ILF, respectively.

6 Measurement Data

The eleven case studies were selected taking into account their quality. The authors conducted a thorough review in search of inconsistencies in definitions, overlapping definitions, and generally poor implementation of the given heuristic [9] to generate the LEL and scenarios. In case of finding any of these problems were corrected and recorded in each case study the amendments and if they were insurmountable, the case was dismissed [18]. This process ensures a set of case studies as to its uniform quality. Importantly, these case studies have not been implemented, thus it inhibits the taking of measurements on an artifact produced in later stages.

Table 1 presents the measures of the LEL and the FP of the scenarios for the case studies. The data is ordered in increasing sense as the functional size (last column).

Table 1. LEL and FP Measurements

Case	SL	SV	SS	ISV	SFP FP[1]
Saving plan	38	15	23	28	66
Reception	39	12	27	50	103
Passport	37	6	31	16	103
Mortgage Loan	31	0	31	0	103
Meeting Scheduler	34	13	21	45	123
Library	35	9	26	9	128
Students System	28	1	27	3	143
Blood Bank	55	10	45	16	160

Gas Station	91	22	69	56	216
Notifications	53	9	44	31	279
LEL&Scenarios	78	21	57	83	330

[1]FP calculated with the procedure SFP [1].

7 Statistical Analysis

This study developed and evaluated different linear regression models with data from the case studies presented in Table 1. The following activities were performed:

1. Establish relations between measures that characterize the LEL and the FP.
2. Describe a model for the estimation of FP.
3. Evaluate the model according to different statistical tests.
4. Apply the model with better statistical values for prediction.

7.1 Evaluation Criteria for the Lineal Regression Model

Regression is a technique commonly used to correlate experimental data. It uses a dependent variable and one or more independent variables to obtain a numerical relationship that evidences the influence of independent variables on a dependent variable of the system. When relationship between variables is not known, a linear relationship is assumed, if it does not fit the data, a polynomial or exponential could be tried [12].

Each proposed linear regression model was evaluated using the following statistical parameters: coefficient of determination, *p-value*, MMRE, pred(0.25) and pred(0.30).

The coefficient of determination (r^2) is a measure of the linear association between the two variables *y* and *x*. Takes values in the range of 0 to 1. A value close to 1 means that there is a high degree of coincidence in the variation between two variables, but not necessarily a cause-effect relationship.

The statistical significance (*p-value*) of a result is an estimated measure of the degree to which it is "true" (in the sense of "representative of the population"). In many areas of research, the *p-level* of 0.01 is customarily treated as a "border-line acceptable" error level.

The Mean Magnitude Relative Error (MMRE) is a measure of the discrepancy between actual and calculated values, is expected to be less than or equal to 0.25 or 0.30.

Another widely used prediction quality indicator is *pred(m)* which is simply the percentage of estimates that are within m% of the actual value, where *m* is the expected value for MMRE. Typically m is set to 25 so the indicator reveals what proportion of estimates are within a tolerance of 25% [4]. Values of *pred(m)* greater than or equal to (1-m) are considered acceptable.

Both MMRE as pred(m) are considered as measures of the accuracy of the model and are widely distributed in the analysis of software metrics [3], [8]. Some authors [8] consider that its value does not measure the goodness of the model but MMRE is a measure of the dispersion of the values and pred(m) is a measure of the kurtosis values of the ratio between actual and predicted values. Despite this, these two observations are included as statistical analysis of data, because along with the other measures extend the scope for characterizing the data.

7.2 Regression Model

Linear, polynomial and exponential regression models were analyzed, using different measures of the LEL as independent variables, individually or grouped. This article presents only those models that were significant by the statistical values obtained.

The following are the different regression models proposed and the statistical parameters calculated[1] to determine acceptance or rejection.

Simple Linear Regression Model with Intercept ($\beta_0 \neq 0$)

Table 2. Values of simple linear regression with intercept

	r^2	*p-value*	MMRE	Pred(0.25)	Pred(0.30)
SL	0,548232	0,009159	0,280151	0,454545	0,636363
SV	0,220281	0,145275	0,399296	0,363636	0,545454
SS	0,587048	0,005959	0,275340	0,454545	0,545454
ISV	0,364971	0,049015	0,387699	0,454545	0,636363

The best values obtained for each of the statistics correspond to the row of *noun symbols*: r^2 = 0.587048, p-value = 0.005959, MMRE = 0.275340, pred (0.25) = 0.454545 and pred (0.30) = 0.545454. However, these values are not within the limits required to be accepted by the model.

Multiple Linear Regression Model with intercept ($\beta_0 \neq 0$). The same measures as in Table 2 were considered, combined in pairs, excluding the combination of *LEL symbols* with *noun symbols* or *verb symbols* with *LEL symbols* because noun symbols and verb symbols are included in *LEL symbols*.

Table 3. Values for Multiple Linear Regression with Intercept

	r^2	*p-value*	MMRE	Pred(0.25)	Pred(0.30)
SS	0,587274	0,028489	0,275667	0,454545	0,545454
SV		0,948773			
ISV	0,637486	0,322229	0,26226	0,454545	0,636363
SS		0,039792			
ISV	0,380853	0,187719	0,395460	0,545454	0,545454
SV		0,662583			
SL	0,560260	0,096179	0,287189	0,545454	0,545454
IVS		0,652420			

[1] Statistica 6.0, StatSoft Inc. was used.

As shown in Table 3 all the values are outside the range established as acceptable. Given these results should be analyzed whether any LEL could be contributing to the population with unusually high or low values (outliers) that alter the trend of the whole. To investigate this possibility we calculated the quartiles and the boxes with their corresponding lower and upper values. Thus it was determined that the cases they bring outliers for each of the populations are: *LEL&Scenarios* in *LEL symbols*, *Mortgage Loan* and *Students System* in *verb symbols*, no cases in *noun symbols* and *impacts of verb symbols*.

In Table 4 are the values of r^2 and *p-value* from the Table 2 and those obtained after removing the *LEL symbols* in *LEL&Scenarios* and *verb symbols* in cases *Mortgage Loan* and *Students System*.

Table 4. Values for Simple Linear Regression with intercept and without outliers

	Sin valores extremos		*Con todos los valores*	
	r^2	*p-value*	r^2	*p-value*
SL	0,397521	0,050669	0,548232	0,009159
SV	0,209131	0,215847	0,220281	0,145275

Observing Table 4 one can conclude that the exclusion of one or more case studies does not represent an improvement in the values obtained for the coefficient of determination for *p-value* and therefore no other statistics were calculated (MMRE and pred(m)). Having established the low (and negative) impact of discarding outliers in the following the measures of all cases will be used.

Simple linear Regression Model without intercept (β_0 = 0). A line without intercept means that the independent variable is zero and therefore the value of FP is zero. From a conceptual point of view, it is logical for the modeling problem because if there are no symbols it means that there is no system and thus its size in FP is zero.

In Table 5 presents the results for a Simple Linear Regression Model, such that the line representing it passes through the origin.

Table 5. Values for Simple Linear Regression without intercept

	r^2	*p-value*	MMRE	Pred(0.25)	Pred(0.30)
SL	0,910549	0.000001	0,273790	0,545454	0,636363
SV	0,757906	0,000229	0,558631	0,363636	0,545454
SS	0,919244	0.000001	0,271543	0,363636	0,545454
ISV	0,752948	0,000254	0,567898	0,181818	0,181818

Compared with the results for the simple regression with intercept (Table 2), the values of the coefficients are significantly better for all cases where the calculations are performed with a model without intercept.

The best cases correspond to *LEL symbols* where r^2 = 0,910549, *p-value* = 0.000001, MMRE = 0,273790, pred(0.25) = 0,545454 y pred(0.30) = 0,636363 and *noun symbols* where r^2 = 0,919244, *p-value* = 0.000001, MMRE = 0,271543, pred(0.25) = 0,363636 y pred(0.30) = 0,545454. Even though most of these values are within acceptable ranges, it can not be the same for pred(0.25) and pred(0.30) for both variables (*LEL symbols* and *noun symbols*).

Multiple Linear Regression Model without intercept (β_0 = 0).

Table 6. Values for Multiple Linear Regression without intercept

	r^2	*p-value*	MMRE	Pred(0.25)	Pred(0.30)
SS	0,919306	0,002165	0,272002	0,545454	0,727272
SV		0,935627			
ISV	0,928850	0,298922	0,257942	0,545454	0,636363
SS		0,001094			
ISV	0,773761	0,447524	0,569936	0,181818	0,181818
SV		0,386578			
SL	0,911897	0,002975	0,280137	0,545454	0,545454
ISV		0,719104			

As in the case of Table 3, in Table 6 measures were combined. The best calculated values correspond to the correlation between the FP and *noun symbols* with *verb symbols* and the FP and *noun symbols* with *impacts of verb symbols*. In both cases the value of *p-value* is very high for the independent variables *verb symbols* (0.935627) and *impacts of verb symbols* (0.298922). These terms are not statistically significant and should be removed from the model, so the model could be simplified to a single variable: *noun symbols*.

Since that the LEL is a natural language description of the UD, is significantly influenced by the style of writing and prior knowledge of the environment of the requirements engineer. The analysis different LEL has revealed that some LEL have a higher percentage of symbols belonging to the category verb in comparison with the total amount of symbols, since they stress the description of the actions that take place in the system, while others describe the objects and people producing a LEL with a large percentage of symbols which are in nouns. For the above and to generalize, it was decided to characterize the LEL using the total amount of symbols.

For the case of considering *LEL symbols* as independent variable, that will be denote as FP (SL), the best statistics are those of the simple linear regression model without intercept. In Table 5 presents the values of the statistics.

The correlation coefficient has a value of 0.910549. This indicates that approximately 91% of the variation of the FP is described by the *LEL symbols*.

The value of *p-value* is 10^{-6} which is much lower than the set 10^{-2} to consider the model acceptable. Therefore, one might argue that there is a statistically significant relationship between variables.

MMRE value is 0.273790; this value is very close to 0.25 as the optimal set. The pred(0.25) is 0.545454 compared with 0.75 expected, although this is far from the desired value, it appears that pred(0.30) compared to the value 0.636363 is 0.70 as acceptable, which means that approximately 64% of predicted values are within 30% of actual values.

For all the foregoing it is concluded that the simple linear regression model without intercept is the model that best represents the experimental data.

Table 7. Estimated Parameters for Simple Linear Regression Model without Intercept

	Parameter Estimates (datos proyecto_sim_impl_definitivo.sta) Sigma-restricted parameterization						
Effect	FP Param.	FP Std.Err	FP t	FP p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	FP Beta (ß)
Simbolos	3,321009	0,329164	10,08923	0,000001	2,587586	4,054431	0,954227

In Table 7 reproduces the output generated by the software that displays the values of the parameters estimated for the simple linear regression model without intercept. From these values the value *FP Param* is obtained, which corresponds to the coefficient that relates the FP with LEL symbols. Therefore, the established model corresponds to the following equation:

$$FP(SL) = 3.321 * SL \ . \quad (1)$$

In Table 8 are the predicted values of the FP when applying the equation (1) to the *LEL symbols* (FPp) and the actual values of the FP calculated using the SFP (SFP FP) procedure.

Table 8. Values for predicted FP and actual FP

Case	SL	FPp[1]	SFP FP
Saving Plan	38	126	66
Reception	39	130	103
Passport	37	123	103
Mortgage Loan	31	103	103
Meeting Scheduler	34	113	123
Library	35	116	128
Students System	28	93	143
Blood Bank	55	183	160
Gas Station	91	302	216
Notifications	53	176	279
LEL&Scenarios	78	259	330

7.3 Prediction of New Observations

The regression model was applied to predict the functional size of three systems which data were not used to define the regression model. These systems are: *Programs Presentation System*, *Trade Management System* y *SFP Tool.* The L&S of the first two were generated after the application has already been implemented. The L&S of *SFP Tool* was produced following the heuristics recommended in a conventional manner and subsequently implemented.

The relationships obtained during the regression analysis are only valid for values of the regressor variable within the range of the original data and should not be used for the purpose of extrapolation [10]. In this case, the model is applicable to systems with a *LEL symbol* number within the range used to describe the model (28, 91).

Table 9. Prediction of the functional size

Case	SL	FPest.	SFP FP	%Error[2]
Program Presentation System	41	136	227	40%
SFP Tool	70	232	226	-3%
Trade Management System	69	229	612	63%

As can be seen (Table 9), with the exception of *SFP Tool*, the error of the estimate is relatively high for the *Program Presentation System* and very significant for the *Trade Management System*. It is pertinent to stress that the case of *SFP Tool* is the only one of the three which followed the regular process of developing the L&S first and then build the system.

8 Conclusions and Future Works

This study evaluated the feasibility of an approach of the functional size from LEL.

The application of statistical techniques on data from the L&S for a set of case studies helped to establish a linear regression model to estimate the functional size of a system from the number of symbols of the LEL. The model was evaluated using different statistical tests and the results were satisfactory, besides having the additional advantage that the number of symbols of the LEL is a direct measure that is readily available. While the values of the statistics obtained are encouraging, due to the little number of cases considered, it is not possible to arrive to a definitely conclusion and it is necessary to extend the experiment. To verify its validity, it must be borne in mind that when working with estimates, regardless of the method used, it is essential to reestimate as soon as more information becomes available.

The authors believe that this issue does not end with this initial proposal. The observations and analysis leaves open a number of valuable issues that deserve to be revised in future studies. These are listed in the following paragraphs.

Systems, rather than been analyzed individually, could be grouped according to some common feature and an additional statistical analysis could be done. Cases could be grouped by the number of *LEL symbols*, the size in FP or the average of the *Normalized Words* [2]

Another issue to consider is to extend the analysis to other measures of the LEL that were not observed in this study. For example the correlation between the FP and some of the direct measures proposed in [2]. For instance, among the immediate direct measures could be considered the *Fan In* and among direct standard measures the average of *Total Normalized Words*.

Finally, adding more case studies could be considered further analysis of the correlation between the FP and the measures of the LEL mentioned above, but discarding the data for the case studies that present unusual high or low values (outliers) for these variables.

Note: An unpublished longer version of this paper is available on request. To obtain a copy please write to the first author.

[2] Calculated as (SFP FP-FPest.)/ SFP FP

References

1. Bertolami, M., Oliveros, A., “SFP: Un Procedimiento de Estimación de Puntos Función de Escenarios”, Proceedings 9th Workshop on Requirements Engineering, Rio de Janeiro, Brasil, 2006, pp. 101-108.
2. Centeno, M.E., Oliveros, A., “Análisis de medidas en la etapa de elicitación de requerimientos”, Proceedings 9th Workshop Iberoamericano de Ingeniería de Requisitos y Ambientes de Software, La Plata, Argentina pp. 405-417, 2006.
3. Dolado J., “A validation of the component-based method for software size estimation”, *IEEE Trans. Software Eng*., vol. 26 (nro. 10), pp. 1006–1021, 2000.
4. Fenton, N., Pfleeger, S., Software metrics *A rigorous and Practical approach*, Second Edition, PWS Publishing Company, 1996.
5. Hadad, G., Kaplan, G., Oliveros, A., Leite, J., “Integración de Escenarios con el Léxico Extendido del Lenguaje en la elicitación de requerimientos. Aplicación a un caso real”, Universidad de Belgrano, 1996.
6. IFPUG, Manual para la Medición de Puntos Función, Versión 4.1.1., AEMES, 2000.
7. International ISO/IEC Standard 14143-1, “Information technology - Software measurement - Functional size, Part 1: Definition of concepts”, 1998.
8. Kitchenham, B.A., Pickard, L.M., MacDonell, S.G., Shepperd, M.J., “What accuracy statistics really measure”, *IEE Proceedings - Software Engineering*, vol. 148, pp. 81-85, 2001.
9. Leite, J., Hadad, G., Doorn, J., Kaplan, G., “A Scenario Construction Process”, *Requirements Engineering Journal*, 2000.
10. Montgomery, Runger, *Probabilidad y Estadística aplicadas a la Ingeniería*, 3ª.Ed., Mc. Graw Hill, 1996.
11. Rule, G., “The Importance of the Size of Software Requirements”, Software Measurement Services, *NASSCOM Conference*, India, 2001.
12. Bethea, R., Duran, B., Boullion, T., Statistical Methods for Engineers and Scientists, Marcel Dekker, Inc., 1995.
13. Hill, P., Software Early Lifecycle – Functional Sizing, ISBSG, Jun. 2006, Vol. 9, No. 2, https://www.softwaretechnews.com/stn_view.php?stn_id=5&article_id=23, consultado en noviembre de 2008.
14. Santillo, L., Conte, M., Meli, R., “Early & Quick Function Point: Sizing More with Less Metrics”, p. 41, *11th IEEE International Software Metrics Symposium (METRICS 2005)*, 2005.
15. Choi, S., Park, S., Sugumaran, V., “Function Point Extraction Method from Goal and Scenario Based Requirements Text”, NLDB 2006, LNCS 3999, Springer-Verlag, pp. 12–24, 2006.
16. Karner, G., Resource Estimation for Objectory Projects, Objective Systems SF AB, 1993.
17. Fetcke, T., Abran, A., Nguyen, T., “Mapping the OO-Jacobson Approach into Function Point Analysis”, *Proceedings TOOLS '97*, Santa Barbara, CA, pp. 192-202, 1997.
18. Centeno, M. E., Pautas para la Construcción del LEL, mayo 2006, 16 pág., inédito. Trabajo realizado en el contexto del proyecto de investigación “Estimaciones de Tamaño en la Etapa Inicial de un Proyecto de Desarrollo de Software”. Universidad Nacional de la Patagonia San Juan Bosco.