Linguistic Analysis of Research Drafts of Undergraduate Students

Samuel González López

Aurelio López-López

Reporte técnico No. CCC-13-001
06 de marzo de 2013

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Linguistic Analysis of Research Drafts of Undergraduate Students

Samuel González López[1], Aurelio López López[2]
Coordinación de Ciencias Computacionales
Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Tonantzintla, Puebla México
[1,2]{sgonzalez,allopez}@inaoep,.mx

Abstract

Academic programs or courses conclude with a thesis or research proposals, elaborated by students in most Educational Institutions on México. In this process, students are advised by a professor who spent time on them. However, the graduation rate is low for many of these Institutions. This paper proposes a method to evaluate linguistically essentials sections of proposal drafts. The intention is to help undergraduate students of two kinds, i.e. Bachelor and Advanced College-level Technician level, in initial drafting, and teachers in the early review process. We propose an assessment at four levels; where the first focuses on the lexicon used by the student in his/her draft, the second level seeks to identify and assess the level of coherence, the third level considers language models intended to identify the particular structure of each element of the proposal, and the last one focuses on identifying answers to methodological questions such as *What will you do?* and *How are you going to do it?,* that characterize objectives section of a proposal draft. This proposal presents the initial results in terms of lexical and global coherence analysis of proposal drafts of students. In preliminary results, we found in a lexical analysis experiment that graduate students showed higher levels of lexical richness that undergraduate student. Also, the lexical sophistication was one of the measures that best differentiates both levels. Moreover, additional results show that the level reached so far by the coherence analyzer is adequate to support the review, taking into account the level of agreement with human reviewers for one section.

Keywords: Method, methodological questions guide, language models, lexical analysis, local and global coherence, proposal drafts.

Index

1. Motivation 2

2. Problem Statement 4

3. Proposed Solution 6

4. Previous work 8

5. Research Questions 15

6. Hypothesis 16

7. General Objective 16

7.1. Specific Objectives 16

8. Methodology 17

9. Contributions 19

10. Preliminary results 20

10.1. Lexical Analysis 20

10.2. Global Coherence Evaluation 27

10.3. Proposed models to evaluate a Draft 34

11. References 41

12. Appendix A 43

13. Appendix B 52

1. Motivation

Knowledge generation is an important feature of developed countries, and knowledge societies are fundamental in the development achieved by these countries. In México, research and development of new knowledge is supported by Research Centers and some Universities, private or public. The level of enrollment of college level students has grown, between 2004 and 2009 the increase was of 26%, but there is a big problem that affects all institutions, the percentage of students that successfully obtained their degree. This graduation indicator is about 48% as reported by ANUIES[1].

Most institutions that offer undergraduate programs in México offer the possibility of concluding the program with the preparation of a thesis. For Bachelor level (BA), the percentage of student that prepared a thesis for the period 2008-2009 was 79.1%, according to ANUIES. In this level, students have different options to conclude their careers, one of them is the development of a thesis. For Advanced College-level Technician degree (TSU), the percentage of graduates in the same period was 39.32%. At this level, students finish their program with the development of a short thesis.

Factors affecting the rate of graduation are diverse. In the study by Martinez et al. [1] they identified some factor of administrative nature and other of academic type. Within the academic factors, students reported absence of advice, difficulty in defining the subject to be developed, and the preparation of the thesis project, to name a few. They also concluded that the development of a thesis is difficult for students, because they do not know with certainty the characteristics of the thesis elements.

The process of developing a thesis begins by outlining a proposal draft or research project, commonly involving the academic advisor and the student. During this process, the advisor spends time reviewing the draft that the student formulates and provides recommendations. This becomes a cycle, ending with a proposal that complies with features that have been established in research methodology books and institutional guidelines. Sometimes this cycle slows down and some of the feedback generated by the academic advisor is focused on the structure of the elements of the proposal draft, for instance, the proposal should include a hypothesis or objective. It is important to note that

[1] National Association of Universities and Institutions of Superior Education in México

each element of a research proposal has its own characteristics and these elements have to be interrelated [2].

The lexicon used by students is a feature of all elements of a proposal draft, which should be considered as a condition to satisfy in the final document that the student delivers.

Another feature to analyze in research proposal drafts is studied in [3] and [4], which describes methods for the evaluation of local and global coherence, an aspect that any proposed thesis must comply. Approaches that have been addressed are at syntactic and semantic level. The first approach characterizes the use of an entity in different syntactic positions and how they are distributed between adjacent sentences, while the semantic approach searches the thematic connection between the sentences. Nevertheless, coherence is only one element of several that advisors review.

The syntactic structure of each element on a research project is another feature that could be handled, i.e. how the students construct their sentences in each of the elements. For example, objectives mostly start with a verb in infinitive, and the research questions follow the structure of an interrogative sentence. These syntactic features of each element become important at the time the students write their research project. Some studies have used language models to characterize the text, mostly in speech recognition which is supported by probabilistic models, to correct certain errors that could generate when transcribing speech to text [5].

Another feature identified in some proposal elements, specifically in objectives and justification, is referred to answer methodological questions that serve as guide for their construction, these questions are *What will you do?, How are you going to do it?.* For the justification element, other question is considered, such as *Who will benefit?.*

These questions involve answers that require to achieve a reflection of the student and the structuring of various terms, i.e. the responses do not fit specific data. Currently the search for answers to questions has been studied to find dates, places or names of people in [6] and [7]. The question-answering technique has been used for the retrieval of specific information and part of a query expressed in natural language. The process performed in this technique includes an analysis of the question, information retrieval and the extraction of answers.

We can say that using natural language techniques could help to provide support in the analysis of a research proposal draft with emphasis on the specific linguistic analysis that each element of a project proposal requires. This work seeks to create methods to help the language assessment of certain characteristics of the elements on a research proposal, as the lexicon, coherence, syntactic structure proper to each of the elements, and identifying answers to methodological questions.

2. Problem Statement

Based on the experience as academic advisors, the first drafts elaborated by students exhibit a lot of deficiencies. It is known beforehand that a proposal draft indicates the first attempt to express an idea in a structured document. This idea usually is not definitive, even the phrase, "*The first draft of anything is shit*" expressed by Ernest Hemingway alludes to the difficulty in the process of writing a proposal, which involves improving the document in further versions.

Initial deficiencies appear at lexical level, for example the repetition of words within a paragraph or the absence of technical terms of the domain. The lexicon is an aspect that students must comply from the beginning, but often is not satisfied, hence the interest in our work to attack this problem as a necessary condition of a proposal. In this way the student will achieve an acceptable level regarding the writing of his/her proposal. Other deficiencies can be at a higher level as the absence of arguments for an idea.

As a result of poor writing, the adviser requires more time to reviewing the structure of the draft and dedicate less time to examining the content. In addition the progress of writing by the student is slow.

Identifying answers to methodological questions involves challenges for computational linguistics, as this does not searches only for specific data, we seek a sequence of terms that respond the questions that allowed the construction of an objective or justification. For instance, the next objective of a research project:

Develop an inductive learning algorithm to solve binary classification problems from unbalanced data sets, where the results reach appropriate consensus between accuracy and comprehensibility.

From the objective statement is possible to identify answers to the questions:

- What will you do? : *Develop an inductive learning algorithm.*
- What for will be done?: S*olve binary classification problems from unbalanced data sets.*

In the answers, we observe the result of a process of reflection that the students have to go through to translate their ideas. Since one does know beforehand the "answer", it would be difficult to identify such answers. Therefore, it requires a new approach. It is worth mentioning that in this level of identifying answers to methodological questions, we seek to support the student from a structural approach, i.e., we do not attempt to understand the content of the answers, we want that the student has reflected in its objective or justification a sequence of terms corresponding to the result of the reflection of the questions stated above.

Some papers on question answering have divided a complex question in simplest questions using connectors that contains the same question, these connectors are identified by a part of speech tagger [8]. This approach does not seem to be enough for our purpose.

Several of the review processes are performed naturally by the academic advisors. For example when considering the structure of justification beforehand, the advisor has established the definition of justification and argument to be established in this section taking these aspects of the review implicitly. Therefore, this structure not only comes down to a combination of syntactic elements or sequence of tokens, involves a deeper analysis by the academic advisor.

Authors of research methodology suggest that a justification has to show evidence that some aspects have emerged, such as importance or needs of the problem. They also mention of the benefits of work, as well as beneficiaries. The academic advisor has to scrutinize for these concepts when going through the justification section. This task represents a challenge for a computational approach, and this research attempts to address.

Other aspect of our interest is reviewing the coherence, that is a requirement that the academic advisor assesses implicitly, and sometimes becomes complex and hence often ignored. Given the complexity of our language, coherence is difficult to analyze, especially when involves the use of pronouns within a paragraph. For an advisor, it may not be difficult to identify that within the paragraph is still talking about the same subject and therefore the paragraph is coherent. This phenomenon of anaphoric elements is difficult to

solve computationally. Our proposal seeks to analyze coherence, to contribute to the overall assessment of a proposal.

The different issues just described imply the incorporation of varied techniques of natural language and in some cases the design of new methods. This work proposes to solve these problems, analyzing and evaluating a proposal draft of a student at different levels, providing support and feedback that cover key initial details.

3. Proposed Solution

Our proposed solution suggests an evaluation at four levels, starting at a basic level as the first filter of the proposed draft, reaching a level of complex assessment (see Figure 1).

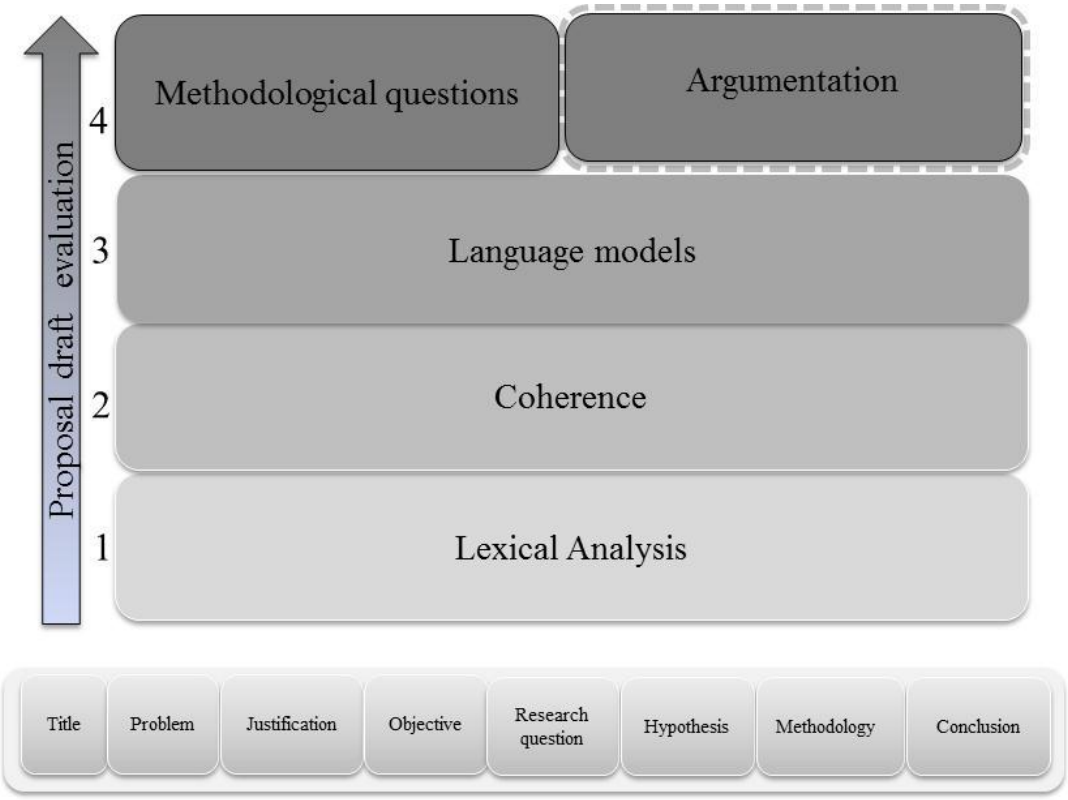

Figure 1: Four- level evaluation

At the bottom of figure 1 are displayed the eight elements of a research proposal draft to be considered as basic elements to evaluate: title, problem statement, justification, objective, research questions, hypothesis, methodology and conclusions. The items will be

treated differently at each level, some will be processed at all four levels and others only in some of them, due to the own characteristics of the elements.

The elements to reach the fourth level are objective and justification, because are those that allow for its construction to answer certain methodological questions. At the first level lexical analysis will be held, this level seeks to determine if the student is using a diverse vocabulary and an adequate balance of content words. Also we aim to determine the richness of vocabulary, based on three dimensions: lexical diversity, lexical density and sophistication For instance, if a student repeatedly writes the word *system* in one of the elements of the project, there will be a reason to suggest that the student has to review the lexicon, trying to reach variety. This is the level at which begins the evaluation and gives a perspective to a basic language level.

The second level focuses on evaluating the coherence locally and globally and is based on the combination of semantic and syntactic approach. This level seeks to capture whether the elements of the proposed research are semantically coherent to the area of computing and information technologies, but also seeks that the elements themselves are coherent, by incorporating the syntactic approach. Combining the two approaches for coherence is considered in previous studies, which showed that the techniques used capture complementary aspects of coherence [9]. At this level, the item title is not pertinent for analysis.

The third level seeks to attack the task of evaluating a proposal draft in language models capturing the syntactic structure that maintains each element, the third level aims for a thematic independence, i.e., we want to capture what kind of elements of discourse are being used and how they are used, such as the use of verbs, adverbs, nouns. This level is above of the thematic aspect and emphasizes in the syntactic elements, to do this looks to characterize each element and defining a syntactic structure or syntactic pattern.

Finally, the fourth level identifies a set of terms that respond to methodological questions through the implementation of a new method, because previous work in question answering is intended for a specific level and factual data. This is the highest level since not only looks for syntactic or lexical features, but is also intended to identify answers to questions that involve a process of reflection on the student, as illustrated above.

The four-level evaluation will provide a global view of the structure of a proposal draft to the student, i.e. the student would have an outcome for each element of the draft and also he/she can receive feedback with suggestions or recommendations to improve the draft.

Therefore, the proposal should show improvements before being submitted to the academic advisor for review. Also, the academic adviser would have more time to focus on the contents of the proposal documents. This work aims to integrate assessments of each of the levels, in such a way that conform a linguistics and structural evaluation platform.

4. Previous work

Many text definitions include coherence as a necessary feature. A formal definition given in the work of Vilarnovo [10], establishes that the coherence of a text is to connect all parts of a text as a whole: the interrelationship of the various elements of the text. Coherence in proposal drafts of students is important because if it is not present in each of the elements, the idea loses all meaning. Different approaches have been addressed by researchers, some techniques have focused on the semantic aspect when seeking to achieve overall coherence evaluation, while other studies have worked the syntactic aspect, as a way to attack the local coherence.

In the study by Foltz, et al. [11], they evaluated the textual coherence using Latent Semantic Analysis (LSA) technique. This paper shows the coherence prediction by analyzing a set of texts statement by statement of four texts, with a 300-dimensional semantic space, which is constructed based on the first 2000 characters of each of the 30,473 articles of the Encyclopedia of American Academic Groliers. After separation of the four individual sentences texts, the vector of each text was calculated as the sum of the weights (each term), subsequently being compared with the next vector, so the cosine of these two vectors showed the semantic relationship or coherence.

One of the discussions in this paper is whether the LSA technique is a model of text-level knowledge of an expert or novice, they say that depends on the training it has received the LSA system in the application domain. This technique focuses on the latent semantic aspect, which would be a relevant aspect to our work.

In the study by Ferreira and Kotz [12], they evaluate the coherence of police news automatically, i.e. given police news written by a journalist, the evaluation system provided the degree of coherence that the news had. In this study, they used the technique of Latent

Semantic Analysis, first compiling a corpus in the police news domain which served to train the analyzer and from that collection, the analyzer measured the coherence of the news.

The expected result was that the coherence analyzer, will be close to the evaluation done by a journalist and a Spanish teacher. The results of six texts of evidence are:

Evaluators	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6
System	0.57	0.72	0.42	0.54	0.71	0.44
Journalist	0.6	0.69	0.54	0.57	0.68	0.48
Spanish teacher	0.66	0.76	0.76	0.7	0.79	0.63

Table 1. Results of the level of coherence in six news

Table 1 shows that the values of level of coherence between the machine and the journalist are close, but between the machine and the Spanish teacher have a considerable difference. Text 3 shows a clear difference in the level values of coherence, the authors explain that was because in the corpus generated for the case of text 3, the word "fire" appeared in a low percentage, i.e. the corpus did not contain sufficient information to assess the text 3. Finally they conclude that the results are positive, since in the newsroom will have a journalist (who may have taken courses in writing) and not a Spanish teacher. The results of this study indicate that the training corpus must be large enough for a good training.

In the research presented by Kulkarni and Caragea [13], they seek to determine the semantic relationship between two words, the first phase uses a concepts extractor, to identify concepts related to the pair of words that are being analyzed and generate its cloud of concepts. In the second phase, the Jaccard coefficient was used to calculate the semantic relationship of the cloud of concepts. The advantage of this work is not restricted to a particular knowledge. A tool that is available on the web[2], allows comparing the similarity of multiple texts in a particular latent semantic space using the LSA technique. It also measures the similarity between adjacent sentences.

For the syntactic approach, a representation of discourse called Entity Grid has been developed, which is built in a two dimensional array, which captures the distribution of entities in discourse between adjacent sentences of text [14]. Each row corresponds to the sentence of the paragraph, and the columns represent the entities of discourse. The array

[2] http://lsa.colorado.edu/

cells contain values that correspond to the roles in the sentence, i.e. subject (S), object (O) or absence of any of the above (X). The main idea of this representation is that while the object and subject are present in paragraph being evaluated, the coherence is strong. They assume that certain types of subject and object transitions indicate that the discourse has a local coherence. Below is an example of this technique.

Sentences	Department	Microsoft	Evidence	Competitors	Markets	Products	Brands	Software	Government	Earnings
1	S	S	X	O	-	-	-	-	-	-
2	-	O	-		X	S	O	O	-	-
3	-	S	O		-	-	-	-	-	-
4	-	S	-		-	-	-	-	-	-
5	-	-	-		-	-	-	-	S	-
6	-	S	-		-	-	-	-	-	O

Table 2. Entity-Grid dimensional array

Table 2 shows entities that were extracted from six sentences, such as the word "Microsoft" in the sentence 1 was labeled as Subject, in sentence 2 was labeled as an object, and in the sentence 3-4 was labeled as a subject, in the sentence 5 this entity was not found, finally in sentence 6 is labeled as a subject.

This transition is shown for the column of Microsoft entity, that reflects a higher density than the other. According to the authors, there are indications that the higher the density of the columns, greater is the coherence level that has the evaluated text. In contrast to the semantic aspect, this technique seeks to capture aspects of local coherence, which is something that our work aims to capture to measure the coherence of the proposal drafts.

In [9], an evaluation of the different techniques used to measure the coherence in text was presented, considering the semantic and syntactic approaches. Techniques that evaluate the semantic approach are based on words, and distributional similarity measures of WordNet, such as, HStO, Lesk, LCON, Lin, Resnik. For the syntactic approach they selected Entity Grid.

In this experiment they used human judgments as the highest level and that was the basis for comparison with each technique. They applied multiple linear regression to determine

the degree of correlation existing between each technique. The results are shown in the following table.

	Humans	Entity Grid	Word-based	LSA	HStO	Lesk	LCon	Lin
Entity Grid	.246							
Word-Based	.120	-.341						
LSA	.230	.042	.013					
HStO	.322	.071	.093	.037				
Lesk	.125	.227	-.032	.098	.380			
LCon	-.290	-.392	.485	.035	.625	.270		
Lin	.173	.074	-.107	.053	.776	.421	.526	
Resnik	.207	-.003	.052	-.063	.746	.410	.606	.809

Table 3. Correlation between human judgment and the different techniques of coherence

Table 3 shows the correlation results, measured with the Pearson coefficient that reached each of the models, with respect to the results of human judgments, where is observed that the technique Latent Semantic Analysis reached .230, the technique Entity Grid correlation reached .246 and the highest correlation obtained was the HStO technique. Observing the correlation between these techniques, it can be seen that they are low, this could indicate that the techniques are capturing different aspects of coherence. One of the lines to work in this proposal is the fusion of some of these techniques, in order to obtain better assessment of coherence.

A combination of algorithm BL08 (that considers nouns and pronouns) for entity grid with writing quality features, such as grammar, word usage, and mechanics errors, showed improvements in the review of the coherence of student's essays on three different populations [15].The experiments used a corpus of 800 essays related to Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE). After performing the experiments, only two out of three populations obtained acceptable Kappa values, between humans and system.

Another component of interest is the syntactic characterization of each element of a proposal draft, through the use of language models, in the studied of Selvan et. al, [16] that presented a lexicalized and statistical parser for word processing of a regional language of India "Tamil". They use language models to generate the probabilities associated with each word and train the models.

They used phrases or dependencies that are in a corpus that has been processed by a parser, where each sentence is represented by a syntactic annotation tree. For the training they used n-grams, where the probability of each word depends on the n-1 word, the best results was reached with n = 3. Then they combined the language model, i.e. the statistical approach with the structural approach, the latter refers to the grammatical structure that has a sentence. Figure 2 shows the parser design.

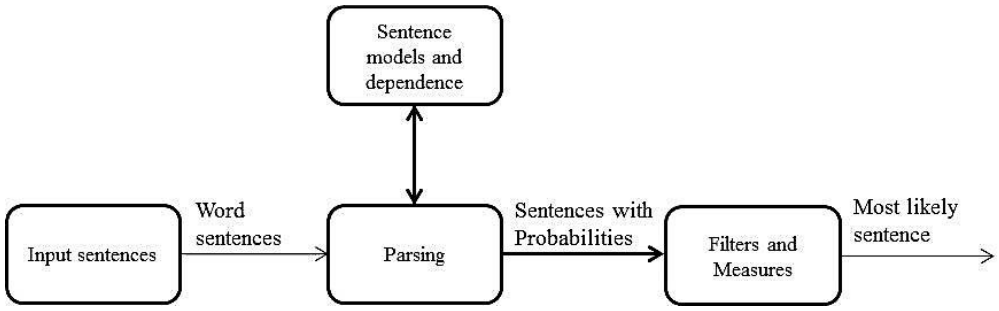

Figure 2. Lexical parser and statistics

Similarly to this paper, our proposal seeks to identify the syntactic patterns in the elements of a proposal draft, through the use of language models. Some works propose the use of language models to generate financial recommendations, specifically seeking financial news stories that might influence the behavior of markets, Lavrenko et al. [17]. This paper proposes a scheme in which two types of information are retrieved to generate the language model, first retrieves information about product prices, which builds with price trends. In parallel, they collect items related to finances, and used a collection of 38,469 articles from Biz Yahoo.

We look for regularities or patterns that could be found in each analyzed element. For example the use of an infinitive verb is a characteristic of an objective. Thus, we found works that have studied the structure of research articles, such as the work about analyzing semi-automatic of structured movements on abstracts of articles, Wu et al. [18]. First they collected abstracts automatically from the Web, which were used for training. Afterwards, each statement in a small sample of 106 abstracts (709 sentences) was manually labeled by four human reviewers, the goal was to create a labeled collection that serves as seeds to train a Markov model. Then, they automatically extracted collocations in order to find phrases that represent rhetorical moves. For example, the collocation “paper address” was

found in the training corpus and was labeled with the type of movement “P”. With this collocation they could tag new sentences.

Another procedure performed in this work was to expand the collocations in order to capture similar movements, but within text not showing the collocation exactly in the same way, for instance, the collocation “address problem” was found in some sentences like “This paper **addresses** the state explosion **problem**”. It is observed that the collocation was expanded. Another example found by the authors was "We **address** the **problem**", and in the same way is observed that the collocation was expanded. Both samples were labeled with the movement "P". The following table shows some examples of collocations found by authors.

Collocation	Move type	Count of colocation with a movement structure	Total of collocation occurrences
We present	P	3,441	3,668
We show	R	1,985	2,069
We propose	P	1,722	1,787
We describe	P	1,505	1,583

Table 4. Example of found collocations

The training corpus contained 20,306 abstracts with 95,960 sentences obtained from Citeseer web site. From the corpus, 72,708 types of collocations were extracted, and only 317 collocations manually with the types of movements. With the Markov model trained, they found a sequence of movements occurring more often: "B-P-M-R-C". This detection model of structure movements could be relevant to our work, but we should consider that we do not have a large corpus, that enabling us to find valid collocations. However, we can consider also the way they use the Markov model.

Other studies have addressed the identification of sections within documents. One is focused on the classification of sections within clinical notes, implementing a HMM. In this paper they used a corpus of clinical notes of New York-Presbyterian Hospital with 9 679 notes. This study considers the sequence and dependence of sections in the notes [19].

The authors identified 15 types of sections: Chief complaint (CC), Assessment and Plan (A/P), Allergies (ALL) Family History (FHX) Social History (SHX), Past Medical History (PMH), Past Surgical History (PSH), Past Medical History and past surgical history (P/P), History of Present Illness (HPI), Laboratory tests (LABS), Physical Examination (PE),

Review of System (ROS), Studies (STUDY), Medication (MEDS), and Health Care Maintenance (H/M).

Thereafter, they trained the HMM with 15 states, each state corresponding to a different section labeled. For a given text window, each observation of each state was modeled using a bigram language model, specific to each section. The aim was to identify each clinical note section as some sections were not labeled (sections headers). The corpus contained 33% of notes without labels section

They also built a dictionary of labels, for example the words "Treatment plan," "impression/plan," and "assessment and plan" were mapped to "A/P". To evaluate the accuracy of the dictionary, they used 120 clinical notes tags for two physicians, reaching 97.36% accuracy. Finally, the corpus was divided in 78% for training and 22% for test, to evaluate the model-classifier. The results of F-Measure were about 90%, statistically above the baseline, which are about 70%. For each note, the accuracy reaches 70% in HMM compared with 19% for the baseline.

Within these results, they found that the section STUDY qualified for error in the LABS section in 10.24%, this was because they are adjacent. Also section CC was classified in A/P with a 23.36%. It can be said that the method is sensitive to adjacent sections. This work could help in one part of our proposal, but emphasizing that we are not seeking to identify the sections of a research proposal, we sought to evaluate the structure of each section.

Within our proposal, at the higher level, there is the argumentation, as an element that is derived from the problem and that could be explored. At this level, we find a work that seeks to identify the organizational level of the elements in an argumentative discourse.

An argumentative discourse is one that presents sentences showing evidence in favor or contrary on a specific theme or when someone wants support a new idea. For instance in the following fragment of text "**There is a possibility that** they were a third kind of bear apart from black and grizzly bears ....". It is noted that this shows something contrary to that established in the document [20]. The authors present two approaches to try to identify the level of argument, the first is a rule-based system and the second is a sequential probabilistic model. The first approach seeks semi-automatically patterns, for example a pattern according to a verb is:

AGREEVERB => Disagree | Agree | concur | …

Thereafter they formed a SHELL, which grouped patterns such as:

SHELL => I [MODAL] [ADVERB] AGREEVERB with the AUTHORNOUN

The second approach is a supervised sequential model based on Conditional Random Fields (CRFs), using a small number of lexical features based on frequency. The two approaches could be useful to identify certain sections of a proposal that are argued, such as the justification section.

From the review of the state of the art, we have found that most of the research has focused on the analysis of the structure of abstracts of scientific papers, some automatically generated. But there are interesting techniques used in these studies, and that somehow could be applied or adapted to our proposal.

Even some authors have combinations of techniques, where the results of natural language techniques have been the entry of other techniques. For example the use of a classifier based on results sent by language models where the perplexity measurement is the input.

Some techniques reviewed in the state of the art, we have applied, such as Latent Semantic Analysis (LSA), this technique was applied to detect the global coherence, using the collected corpus. It is worth mentioning that this technique has been applied in other documents and domains.

Finally, is necessary to continue the review of the state of the art to achieve generate an appropriate method that allows us to reach the evaluation proposed in our research.

5. Research Questions

From the discussion in the previous sections, the following research questions emerged, that this research aims to answer:

- *How will natural language techniques help to assess the main sections of a research proposal draft, considering some features that institutional guidelines and authors of research methodology have established?*

- *How to identify the answers to methodological questions such as: What to will you do?, What for will be done?, How are you going to do it?, Who will benefit? in objective and justification elements, to help students improve his/her writings?*

- *How to merge the semantic and syntactic approaches to improve the assessment of coherence when reviewing the elements of a research proposal draft?*

- *What configurations of language models can provide better support to the student and improve the syntax in the different sections of a draft?*

- *How can natural language processing techniques be applied to automatically evaluate the essentials features within a proposal draft that research methodology authors suggest ?*

6. Hypothesis

The analysis and evaluation at four-levels allows the assessment of main features on elements of a research proposal draft, which can provide students with feedback in early stages of its development.

7. General Objective

Design a method to integrate different levels of assessment to analyze linguistically proposal drafts of students, ranging from Advanced College-level Technician degree to Bachelor level, using techniques from natural language processing, reaching acceptable levels compared to human reviewers.

7.1. Specific Objectives

- Design an evaluation method to analyze the vocabulary of each elements on a proposal draft.
- Design a coherence analyzer, incorporating semantic and syntactic approaches, which allow evaluation of the elements and a global perspective of the proposal draft.
- Build language models to characterize each element of a research project, allowing to generate a syntactic pattern.
- Define a method to identify answers to methodological questions within the objective and justification elements, to help identify that there exists answer to

questions, such as, What to will you do?, What for will be done?, How are you going to do it?, Who will benefit?.

Experimentally validate the results generated in each of the levels that are proposed to evaluate a research proposal draft.

8. Methodology

To achieve the objectives planned, we propose to carry out an assessment at four levels: the first level refers to the assessment of lexicon, the second level focuses on the evaluation of local and global coherence, the third level refers to the language models that characterize the elements of a project and the highest level focuses on identifying answers to methodological questions. Each level demands specific techniques, given the varied nature of features to analyze.

1. To describe the general problem, it will be necessary to select the elements of a proposal draft to focus on, considering the computational viability. For this purpose, we will conduct a review of several books of research methodology.
2. Gather a corpus, considering proposals for research projects and theses that would allow to identify features of interest. Proposals for research projects will be gathered considering that are written in Spanish. The corpus will be also useful for carry out experiments.
3. A lexical analyzer for each of the elements on a proposal draft will be designed and implemented, using the following procedure:
    a. Adapt the corpus for lexical richness of each element.
    b. Explore different techniques to measure lexical richness.
    c. Compute the lexical richness of each element on a proposal draft
    d. Compare the lexical measures obtained from each element of a proposal draft, respect to the elements obtained from the corpus.
    e. Define a scale to measure lexical richness.
4. A Coherence analyzer will be designed and implemented, using the corpus. This analyzer will have the following stages:
    a. Adapt the corpus for the evaluation of each element of a proposal draft.

b. Implement technical Latent Semantic Analysis (LSA) + Entity Grid to assess the coherence from the syntactic and semantic approaches.
c. Perform cross-validation to evaluate the results produced by the analyzer.
d. Generate a gold standard using human reviewers
e. Specify a scale to measure the coherence level.

5. Building language models to characterize each element of a proposal draft. The generation of these models have the following stages:
    a. Adapt the corpus for training.
    b. Determine the technique to build models of each element of a proposal draft.
    c. Perform cross-validation to evaluate the results produced by the analyzer.
    d. Build a scale to measure closeness level between the generated models and models of each element analyzed.
6. Develop a method to identify answers to methodological questions within the target elements, i.e. objectives and justification. We foresee the following steps:
    a. Adapt the corpus for processing on the different proposed strategies.
    b. Identify elements of syntax and parts of speech that characterize the answers of methodological questions.
    c. Explore moves of structure in the answers to methodological questions, for instance in this text segment an objective “Develop an inductive learning algorithm”, would represent the move corresponding to What [18].
    d. Develop a method to carry the objective and justification to a higher structural level, i.e. the idea is that method will be independent of the content and allows identifying the answers to methodological questions guide.
    e. Determine that the objective and the justification element, show evidence that student has taken care of the methodological questions
    f. Design an evaluation methodology to validate the method.
7. Experimentally validated the results generated in each of the levels proposed for evaluating a proposal draft. The following stages are defined:
    a. Initialization stage:
        a.1. Select a sample of research proposal drafts on undergraduate level.

a.2. Select human reviewers experienced as academic advisor to give the level of acceptance of the proposal or elements, considering:

- The lexical aspect
- Local and Global coherence
- The syntax of text
- Identification of the answers to methodological questions

a.3. Select a human reviewer to perform the comparison of results obtained by humans and the system.

b. Development stage

b.1. Design a series of steps for human reviewers to evaluate manually the students proposals, taking into account the parameters defined or aspects of interest.

b.2. Evaluate the draft with the computational tool generated, based on the evaluation at 4 levels proposed in this proposal.

c. Analysis stage

c.1. Compare the results of human reviewers with our four-level model. This task is performed by a human reviewer.

c.2. Performing a statistical analysis of agreement among human reviewers. Afterwards another analysis of agreement between humans and our model will be performed.

9. Contributions

Based on the objectives, we expect to contribute the following:

- Create a method to evaluate a student's proposal draft. With this method we could support students and teachers involved in the development of a proposal.
- A lexical analyzer to intended to improve student writing in terms of vocabulary, based on diversity, lexical density and sophistication.
- An analyzer that captures the semantic and syntactic aspects of coherence in the domain of computers science and information technologies.
- Language models specific for main sections of a proposal draft.

- A method to identify answers to methodological questions. This method could significantly assist students in building their objectives and justification.
- A method to integrate the four levels of evaluation properly, so that together they can evaluate a proposal draft, providing feedback to the student.
- A web implementation to evaluate proposal draft of students in early stages of preparation. This implementation would help in the validation stages of our general model.

10. Preliminary results

Several experiments have been developed, these were focused on the first two levels of evaluation of the general model: Lexical Analysis and Global Coherence evaluation. The experiments are according to the specific objectives for research. Some advances on languages models are also discussed.

10.1. Lexical Analysis

The first experiment had as purpose to evaluate the lexical richness of seven sections of a research proposal draft. This analysis is the starting point of the evaluation model proposed, the solution will be the initial stage for the student. We foresee that reaching a medium or high level in the evaluated draft will allow to continue the evaluation with the rest of the levels proposed in the model.

Objective of the experiment

Design a lexical analyzer, considering for the lexical richness of seven sections of a proposal draft, considering the lexical diversity, lexical density and lexical sophistication.

S*pecific objectives*

1. Compare the levels of lexical richness between Graduate students and Undergraduate students (BA and TSU).
2. Perform a correlation analysis to detect possible dependencies between the three measures.
3. Generate a rating scale and design a Web interface to the Lexical Analyzer.

Methodology of experiment

We gathered a corpus of the different elements within proposal documents in Spanish. We distinguished in this corpus two kinds of student texts: graduate proposal documents, and undergraduate drafts. The whole corpus consists of a total of 410 collected samples, as detailed in Table 5.

Corpus		
Section	Graduate	Undergraduate
Problem statement	40	14
Justification	40	18
Research Questions	40	10
Hypothesis	40	20
Objectives	60	20
Methodology	40	14
Conclusions	40	14

Table 5. Spanish Text Corpus

Dimension descriptions		
Dimension	Labels	Formulates
Variety	LV	*Tlex/Nlex*
Density	LD	*Tlex/N*
Sophistication	LS	*NSlex/Nlex*
Tlex: Unique lexical terms		
Nlex: Total lexical terms		
Nslex: Words out of a list of common words(SRA)		
N: Total tokens		

Table 6. Measures to compute lexical richness

The first kind of texts includes documents of master and doctoral degree (PG). The second kind includes documents of Bachelor (BA) and Advanced College-level Technician degree (TSU). The corpus domain is computing and information technologies.

To evaluate the seven elements contained in a research proposal, we proposed a computational model that will include three lexical dimensions. The first step in the model considers the preprocessing of each element. Each section in this module is processed with the Freeling[3] tool to obtain the word stems, converting the analyzed word in its singular, grouping similar terms, and allowing speeding of the lexical analysis (see Figure 8).

Another step in the preprocessing of the text was filtering and removing empty words from a list of 209 words provided for NLTK (Natural Language Tool Kit). Stop words include prepositions, conjunctions, articles and pronouns. After this step, only remained content words, which allowed the calculation of the three dimensions.

In the evaluation module of sections, we define three methods for the calculation of the dimensions. The first procedure is the lexical variety which seeks to measure student ability to write their ideas with a varied vocabulary. This function is calculated by dividing the unique lexical types (*Tlex*) between all lexical types (*Nlex*). The second module refers to

[3] This software is available at http://nlp.lsi.upc.edu/.

the computing of the lexical density, whose goal is to reflect the proportion of content words after removing empty words.

This dimension is obtained by dividing the unique lexical types or content words (*Tlex*) between total words of evaluated text (*N*) i.e. the number of words before removing stop words (see Table 6).

Finally, the sophistication method reveals the knowledge of the technical subject and is the proportion of "advanced" or "sophisticated" words employed. This measure is computed as the percentage of words out of a list of common words (in our case, the 1000 common words, according to Spanish Royal Academy).

Each of the measures takes values between 0 and 1, where 1 indicates a high lexical value, and values close to zero mean a low value of the lexicon of the evaluated section. Together, the three dimensions will identify the lexical richness level of student writing. The sophistication would be a plus for the undergraduate student.

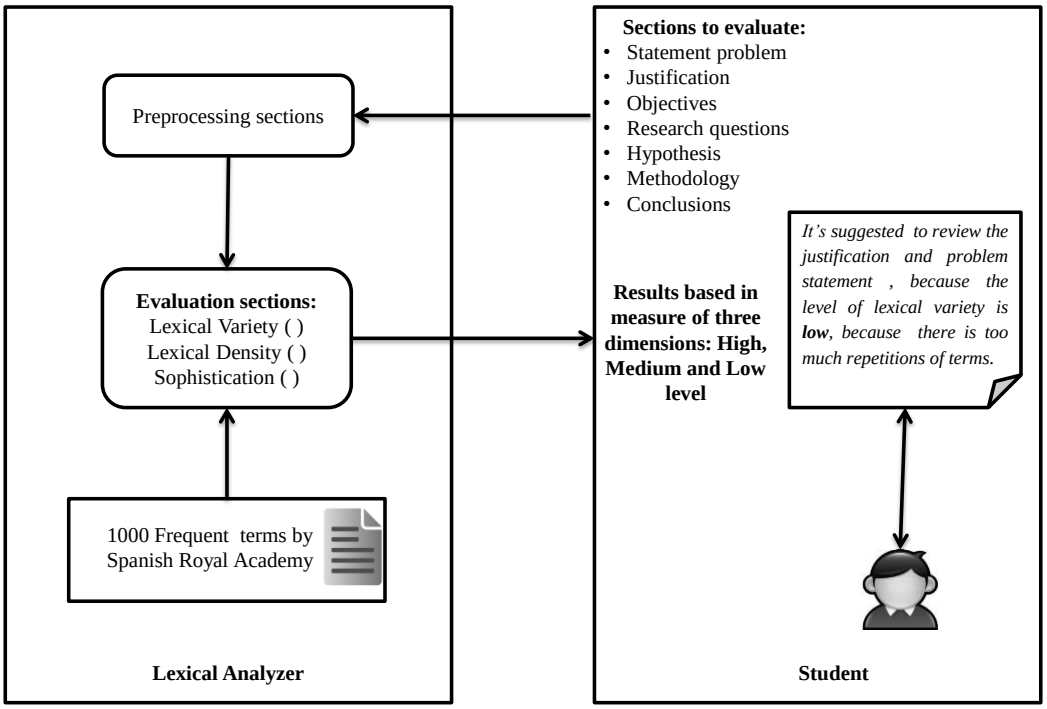

Figure 8 . Lexical Richness Evaluation Model

Both levels (graduate and undergraduate) were evaluated considering the three dimensions in order to make a comparison of lexical richness among them. A correlation analysis between the dimensions was also performed to detect possible relations of dependence. The results provide a guideline to be used as a corpus graduate reference and

to establish a scale to evaluate new undergraduate level drafts. For each section, a scale with following levels was established: Low, Medium and High lexical richness. The High level is defined as one standard deviation (Sigma) above average, Low as one standard deviation below average, and Medium, in between. Consequently, we obtained different ranges that define the scale for the seven sections of a draft. Finally, a web interface was designed so that students can be evaluated based on the three dimensions of the proposal draft and improve it if a result of low lexical richness is obtained (see Figure 10).

Development and results of the experiment

The results were divided into two groups, considering the extent of the sections. The first block groups the sections with short texts and the second with long texts. Research questions and hypotheses for the TSU degree level do not appear because they are not usually required in the proposal drafts. In the first block, documents of graduate level scored better on all three dimensions (see Table 7, lexical richness).

Sections	Lexical Richness			Correlations		
	LV	LD	LS	LV - LD	LV - LS	LD - LS
Objectives PG	0,9187	0,6266	0,6556	0,0659	0,1482	-0,0383
Objectives BA	0,8983	0,5876	0,5878	0,0882	0,3439	0,2296
Objectives TSU	0,8645	0,5654	0,5453	0,8318	0,7995	0,4000
Questions PG	0,9508	0,6743	0,6754	-0,0979	0,4374	0,0596
Questions BA	0,9473	0,5902	0,6356	-0,1785	0,5725	-0,2558
Hypothesis PG	0,9368	0,5919	0,6189	-0,2391	0,2014	0,1153
Hypothesis BA	0,9184	0,5476	0,5968	-0,3907	0,3448	-0,1239

Table 7. Lexical richness and correlations: first block

Regarding the objectives section, TSU degree obtained the lowest value. These results confirm that, as expected, graduate students have better skills in writing research proposals when compared to the undergraduate level in the three sections of block 1.

When performing a correlation analysis between the dimensions of each evaluated level, we found evidence of a correlation arising from the undergraduate and graduate level between LV and LS; this means that the appearance of more content words probably are

reflected as sophisticated words in the text. The lexical density and variety do not show dependence because the correlation was low (see Table 7, correlations).

It was observed in the objectives section of TSU documents that the correlation is strong between LV-LD and LV-LS, allowing the interpretation that an increase or decrease in some of the dimensions, affects the other dimensions. Considering the measures of lexical richness in this section of the corpus, the results of correlated dimensions caused a low level of lexical richness.

In the second block, the results show a slight variation from the first block. We see that the lexical richness values that are more distant in the two levels corresponding to the sophistication dimension, the graduate level being the highest (see Table 8). This dimension could be used to differentiate the levels. Moreover the graduate level could be used for reference, since it showed higher sophistication.

Sections	Lexical Richness			Correlations		
	LV	LD	LS	LV - LD	LV - LS	LD - LS
Problem PG	0,6409	0,5939	0,603	-0,1854	0,3247	-0,0549
Problem BA	0,6441	0,5889	0,549	-0,0407	0,1636	0,545
Problem TSU	0,609	0,5292	0,443	0,0568	0,0568	0,9955
Justification PG	0,6789	0,568	0,583	0,0997	-0,0612	0,1982
Justification BA	0,6679	0,5389	0,523	0,1916	-0,1734	-0,3251
Justification TSU	0,6407	0,5507	0,463	-0,5554	0,9547	-0,7778
Methodology PG	0,6508	0,5838	0,637	-0,2396	-0,1273	0,111
Methodology BA	0,5846	0,5715	0,586	0,1599	-0,0335	0,2738
Methodology TSU	0,6019	0,5589	0,546	0,4734	0,8918	0,7709
Conclusions PG	0,6477	0,5843	0,606	0,1998	-0,0986	-0,1574
Conclusions BA	0,6582	0,5608	0,549	0,5792	-0,5258	-0,4881
Conclusions TSU	0,6612	0,5714	0,469	0,5454	-0,9732	-0,4157

Table 8. Lexical richness and correlations: second block

In the conclusions section, lexical richness in undergraduate achieves the best result, showing that students at this level have better skills to draw conclusions. Nonetheless, reviewing the conclusions of graduate level, it was detected that the number of terms was twice of that of the average undergraduate level (BA and TSU). When performing correlation analysis on the second block sections, negative values were observed for

justification and conclusion sections. The negative correlation means that as a dimension increases, the other decreases.

In this case, the variety dimension is higher than the density. Also, considering the sizes of the texts and having lower density is likely to find higher lexical variety and surmounting the graduate level. Furthermore, these two sections are observed to be shorter than expected for a conclusion or justification since they are as long as a paragraph, in which do not present sufficient arguments, as expected or suggested by the authors in methodology. Therefore, the undergraduate level could not be considered better than graduate level.

Finally, Figure 4 depicts the average of the three measures obtained for both subsets of the corpus and Figure 9 shows the Web interface of lexical analysis. As expected, one can notice that graduate documents produced higher averages than undergrad drafts, for the different elements (sections).

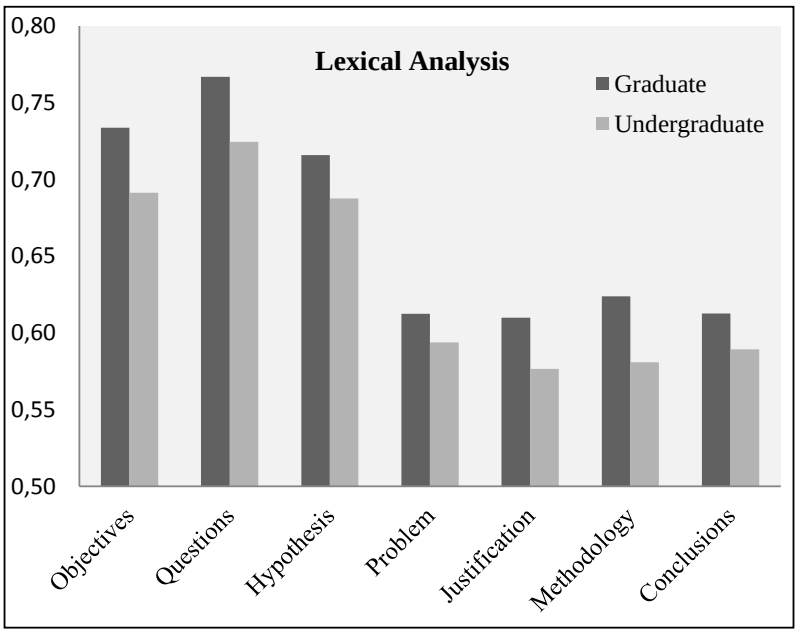

Figure 9. Lexical analysis

Applying the previously defined scale, we obtained the following examples of objectives with High and Low marks in their lexical analysis:

Objective (High level*): Implement an algorithm based on hierarchical structures with enveloping volume of spheres for collision detection.*

Objective (Low level): *Create an information management system for franchises with relevant data of each establishment and personnel data of each franchise, as well as references of franchisees and personal of trust that manage the franchises.*

One can notice that the first example is succinct and concrete, whereas the second example is quite verbose, with scarce technical terms, and abusing of the term *franchise*, lowering its lexical diversity.

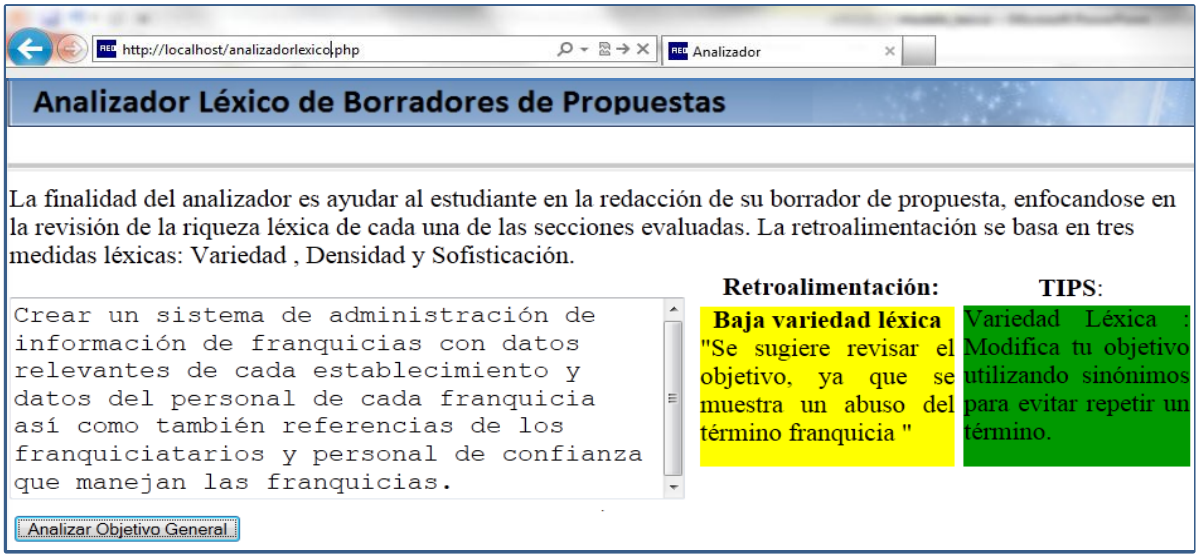

Figure 10. Web Interface of Lexical Analyzer

Conclusions

From our corpus, we can state that graduate students in the area of information technologies have better writing skills. These results allowed us to develop a web tool where students can analyze the vocabulary of each of the essential sections of their proposal drafts.

An interesting finding was that the dimension that differentiates between them more clearly was the sophistication, where graduate levels showed infrequent vocabulary terms. Another aspect observed was the high values of lexical richness in three dimensions obtained by undergrad students for the second block. This result does not imply a

successful result because it would be necessary to verify that the student really does argue properly in each of the sections of a proposal draft, as suggested by the authors of research methodology.

We will continue increasing the size of the corpus, so that the lexical richness analyzer has a higher coverage, since the computing and information technologies domain is quiet extensive and growing. We hope that this computational tool motivates students to develop their drafts and this analyzer contributes to their advancement.

The results obtained of this experiment are reported in a paper that was submitted to the XXV Congreso Nacional y XI Congreso Internacional de Informática y Computación CNCIIC-ANIEI 2012.

10.2. Global Coherence Evaluation

This experiment focused on evaluating the sections of student's proposal draft from the aspect of global coherence. In related work the LSA technique has been used for collections the other domain as police news. In our case this technique was selected because it focuses to capture the latent semantic of different sections in a proposal draft.

Objective of the experiment

Design a Global Coherence analyzer for seven sections of a draft proposal, using Latent Semantic Analysis technique, to generate a scale of coherence that allow identify a coherence level.

S*pecific objectives*

1. Generate a coherence analyzer for the objective section
2. Perform an agreement analysis between human reviewers and our coherence analyzer.
3. Generate a rating scale and design a Web interface to the Coherence Analyzer.

Methodology of experiment

We gathered a corpus of the different elements in proposal documents in Spanish. We

distinguished in this corpus two kinds of student texts: graduate proposal documents, and undergraduate drafts. The first kind of texts includes documents already reviewed and approved by faculty, so they are considered as reference or training examples. The second kind of documents are used as test examples. The whole corpus consists of a total of 410 collected samples as detailed in Table 5, and for the objective section, 60 were for training and 20 were for test. The corpus domain is computing and information technologies. They were then processed removing stop words and applying lemmatization using the tool Freeling.

To assess global coherence using Latent Semantic Analysis in drafts, we set an experiment to validate our process (see Figure 11). First, we asked three instructors to evaluate our whole collection of objectives (training and test subsets), eighty in total. We evaluate the level of agreement among evaluators. Then, we computed the semantic spaces for the different sections of our training subset and then evaluating automatically the objectives in the test subset. Finally, we evaluate the level of agreement between the grade assigned by the system and by instructors. Also we developed an interface web to students. This interface web allowed paste the section to evaluate and shows the result in range: Low, Medium and High Coherence. This scale was created with the cross-validation using the objectives that human reviewers assigned a high level of coherence (i.e. 40 objectives).

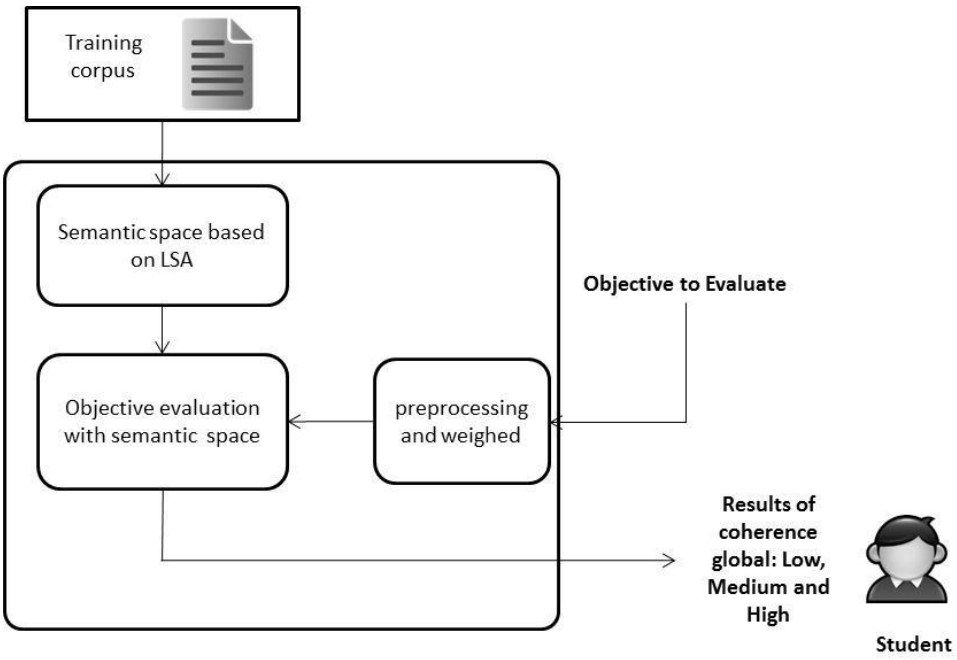

Figure 11. Coherence evaluation model

Development and results of the experiment

All the collection of objectives was sent for evaluation to three instructors serving as reviewers, that have experience in advising students in the preparation of their drafts in the computing and information technologies. The reviewers did not know beforehand the level (graduate or undergraduate) of each sample. Each reviewer was requested to assign a level to each sample, using the scale: High, Medium and Low coherence, where the highest level meant that the text has a strong coherence or relationship to the domain of computing, and the low level means that the relationship is weak relative to the domain. Two examples of High and Low coherence in the objectives section are given next.

High Coherence: *Analyze problems that arise in the system development of software architectures of Enterprise type.*

We can observe that the word "systems", and "software" are very close to the domain, including the term "architecture" surrounded by the above terms fit within the domain of computing. Likewise, words with less thematic load such as "development" or "analyze" are close to the domain.

Low Coherence: *Identify the effect of feedback on the learning of the business leader, to allow to be more effective.*

Notice that even though terms like "learning" or "feedback" may have some proximity to the domain, the words or phrases "business", "leader" or "be more effective" are the central topic and do not match the domain of interest.

The assessments provided by our reviewers allowed to exclude those examples in our training set considered low by at least two, or those where they did not agree, since they will bias the construction of the semantic spaces. On the other hand, the assessments on the test set allow comparing the automatic evaluation of coherence.

Figure 12 shows the percentages of level coherence assignment on each human reviewer. It is noted that the first and second human reviewer similar percentages obtained in each of the levels assigned.

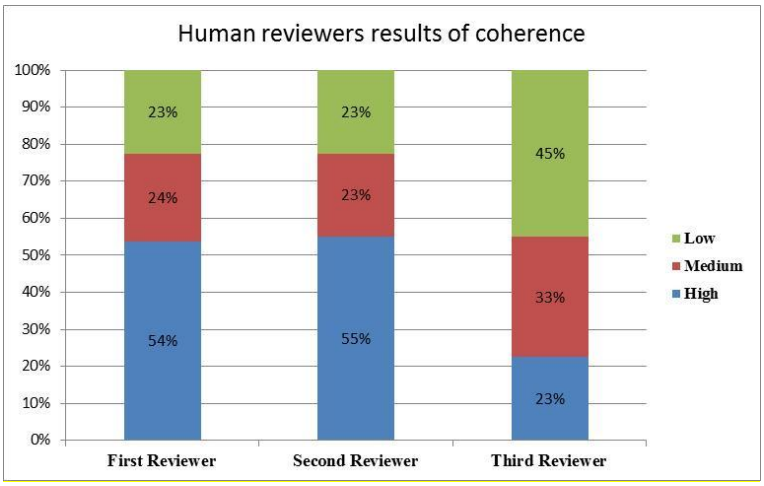

Figure 12. Results of three human reviewers

The third evaluator presented an inverse behavior to the first two reviewers, we can assume that the third rater was more strict when evaluating objectives.

The Fleiss Kappa[4] coefficient of agreement was computed for the three reviewers considering the test corpus. Table 9 shows the Fleiss Kappa results for each level, for the objective section.

Kappa	**Fleiss**	**Cohen**
	Reviewers	Coherence analyzer
High	0,6862	0,0000
Medium	-0,0378	0,2609

[4] Kappa(Landis y Koch, 1977)

Low	0,7353	0,4218
Overall	0,5458	0,2237

Table 9. Kappa for test corpus

The reviewers had a Substantial agreement for the Low and High grading, and a Poor agreement in the Medium. For the results obtained, we conclude that reviewers clearly identified High and Low levels. The overall level achieved between evaluators was 0.54, this giving Moderate confidence of agreement for the experiment.

We built the latent semantic space using the LSA technique with training samples of the whole collection, and in particular of objectives. The texts were processed by removing stop words and applying lemmatization using Freeling.

Similarly as for the lexical analysis, we obtained the scale for automatically assessing coherence in each section, but here we used cross-validation, obtaining the levels High, Medium and Low, according to ranges of values produced by LSA.

These levels allow automating the evaluation of the coherence analyzer. In particular, for the objective section, we got an average of 0.49 with a standard deviation of 0.17, resulting in the highest threshold of 0.64 and the lowest threshold at 0.28.

Once the scale is defined, we evaluated the test samples with the aim to compare the results produced by human evaluators. In this case, Cohen's Kappa is pertinent to compare the level of agreement between human and our coherence analyzer results. Table 2 shows the Cohen's Kappa results for the human versus coherence analyzer.

We observed that the levels of agreement in the Low case is Moderate and Medium level is Fair, the overall level of agreement between humans and the analyzer was Fair. We conclude that the analyzer would have an acceptable support for the student and academic advisor in the process of preparing the proposal draft.

After comparing the statistical results, in terms of the Kappa coefficient of agreement, we also performed a qualitative analysis between the results of coherence analyzer and the process of reviewing a proposal draft, i.e. the advisor would expect that the analyzer was a first filter so that when the drafts reach him, at least have a Medium or High Level. Under this premise, the results of our analyzer match the concept of a strict filtering reviewer, because it provided low and medium values in most test objectives.

We can observe that if our system does not achieve at this time a higher level of agreement in the high level, this is not a problem since the analyzer is being stricter to assign the high level.

In the experiment, the analyzer evaluated as Medium the few highest levels assigned by the reviewers. If the analyzer behaves more flexible and allows high level to objectives that have to be of a medium or low level, this could cause a burden to the academic advisor, failing to support in review.

Finally we note that between the coherence analyzer and human evaluators, the agreement is Moderate for low levels, bringing confidence that the analyzer is identifying those objectives that were classified as Low level for evaluators.

After assessing coherence, the analyzer sends feedback to the student for the seven selected sections in the draft, and updates the parameters of performance in the intelligent tutor.

The analyzer can send recommendations. For example if the analyzer identifies a low level, the recommendation could be "It's suggested to restructure your objectives, making explicit the technical details of the computing and information technologies domain". The following figure shows an example of an evaluated objective, which achieved a high degree of coherence.

The interface is on the first development stage but it is expected that students can evaluate all sections simultaneously or individually. In addition will be counted with an intelligent tutor to give feedback to the student (see Figure 13).

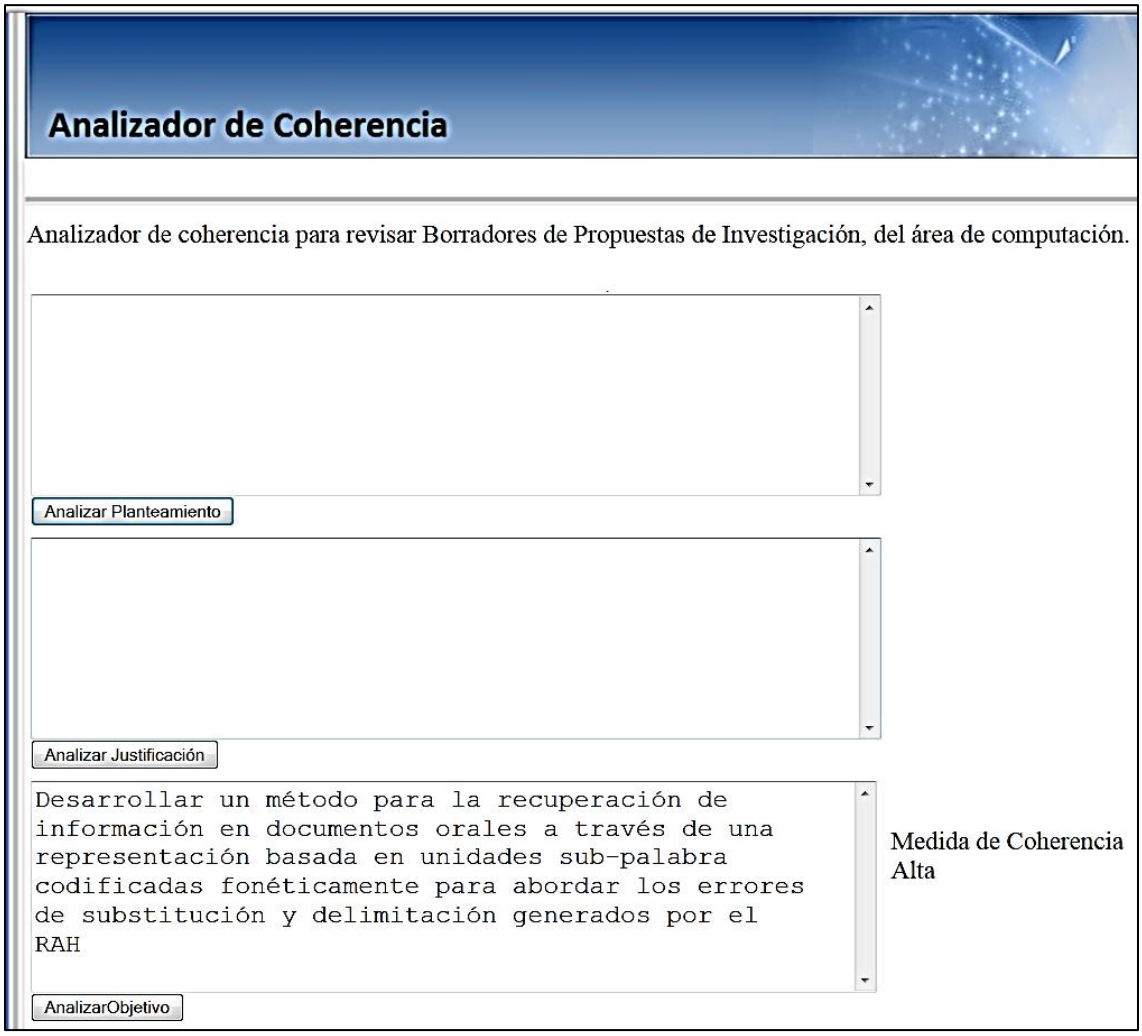

Figure 13: Web Interface of Coherence Analyzer

Conclusions

The LSA technique allowed evaluating the global coherence of objective section in proposal drafts, reaching an acceptable result of the percentage of agreement respect to human reviewers. It was crucial to have a gold standard to compare our results.

We will continue increasing the size of the corpus, so that the analyzer has a higher coverage, since the computing and information technologies domain is quiet extensive and growing.

In these initial experiments, the evaluation of coherence analysis was important to identify the student level, but could be improved by using all the four levels proposed for the evaluation. This will help students to improve their writing, and academic adviser would have more time to review the contents of the proposal documents. This model

assumes that students have prior knowledge to write general documents, so it is not our interest to consider a grammar reviewer.

We expect that this computational tool generates in students a motivation to develop their proposal drafts and this analyzer will contribute to the advance in their proposal drafts. We currently have a web interface for the student to evaluate the draft in the first level, coherence analysis. Bringing our model to a different domain does not seem too challenging, neither moving it to a different language, assuming similar language processing resources and corpus are available.

Currently, we are working on integrating the Entity Grid technique to evaluate the local coherence. We are also exploring the language models based on n-grams for characterizing each element. Also we are in the process of developing a method to identify answers to methodological questions within the elements and objective justification of a proposal draft. We foresee an experiment that includes a pilot test with a control and experimental group of students.

The results obtained of this experiment were reported in a paper and was sent at *The 13th Annual Conference on Information Technology Education and The 1st Annual Conference on Research in Information Technology, SIGITE 2012 & RIIT 2012.* The article was accepted and was presented in October of this year. SIGITE is the ACM Special Interest Group on Information Technology and Education.

10.3. Proposed models to evaluate a Draft

Some proposed models to evaluate each of the elements of a proposal draft are detailed below. The following model is proposed for the evaluation of **objectives, research questions and objectives section.**

We propose the use of language models using the token and the grammatical class of the text to assess, as observed in previous experiments, these sections present some regularity in the writing (see Figure 3).

The goal is to find these regularities and generate a model to compare a new text against the model. In our proposal, the model is trained from a specific corpus, contains examples of objectives, research questions, or justifications.

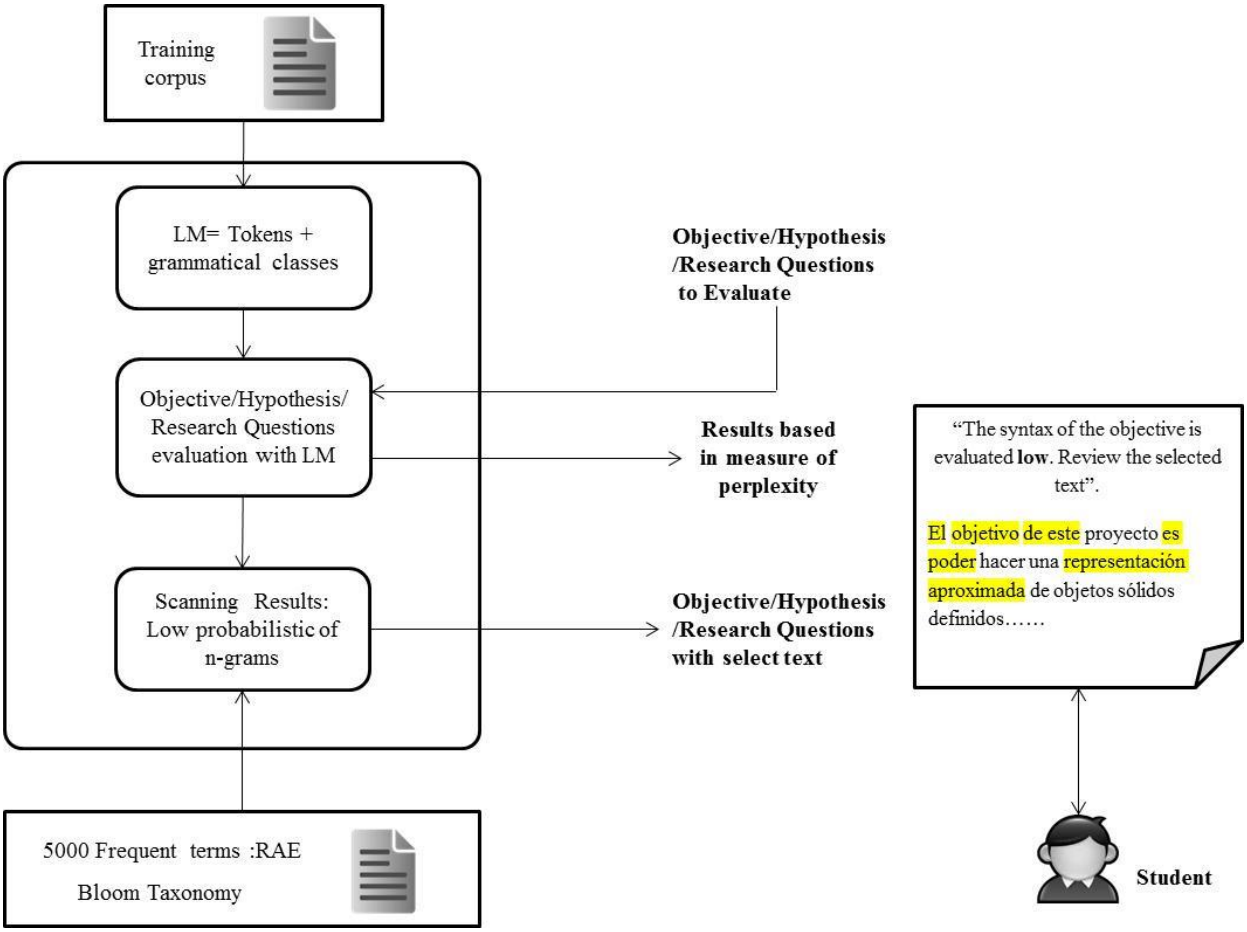

Figure 3. Model to evaluate three elements (sections) of a proposal draft

Later this model will allow evaluate a new text, to obtain a perplexity measure. When the value of perplexity is low, it means that the model could predict properly the evaluated text. Below is the equation for calculating the perplexity:

$$ppl = 10\text{^}(-logprob\ /\ (words - OOVs + sentences))$$

$ppl = perplexity$
$words = total\ terms\ processed$
$sentences = total\ assessed\ sentences$
$OOVs = Total\ of\ terms\ not\ found\ in\ the\ model$
$logprob = probability\ assigned\ by\ the\ model, skipping\ OOVs$

With the perplexity results of model, after running a cross validation will generate a scale that will allow classify the new text. Also we make use of external resources such as more frequent words of the Royal Spanish Academy and Bloom's taxonomy. This taxonomy is a classification of educational objectives, taking in consideration three aspects:

cognitive, affective and psychomotor. In our model, we focus on the cognitive aspect which is subdivided into six levels. Each level is associated with verbs that can be used. For example the verb “describe” is on the first level, corresponding to the level of knowledge, while the verb “analyze” is associated to level four, which corresponds to the level of analysis.

The aim is to use this taxonomy to identify verbs that result from low probability scan. If the verb found corresponds to a low level in Bloom’s Taxonomy it would suggest the student use a new verb. For instance a verb as “leer” (read) or “hacer” (do) would be low level for a research paper, in fact “read” is an implicit activity within a research project.

The frequent terms will be used to identify common words or phrases, such as “esto no nos va” (this does not fit) and suggest the student a change of terms. In the case that the terms are outside the common word list, these terms may be taken as acceptable.

We plan to review the element **Justification** according to the following model (see Figure 4). The model is based on the definitions that authors of research methodology suggest. A justification seeks to explain why the research is relevant, under this premise, five concepts that a justification should include were collected: Importance, Necessity, Benefits, Beneficiaries and Convenience. Inside of justification these concepts can appear as similar, related, synonyms or antonyms terms.

For example, the concept “necessity” may appear in the justification section with the terms “required”, “necessary”, “is required”, “insufficient”. These concepts have been established as dimensions (see Figure 5)

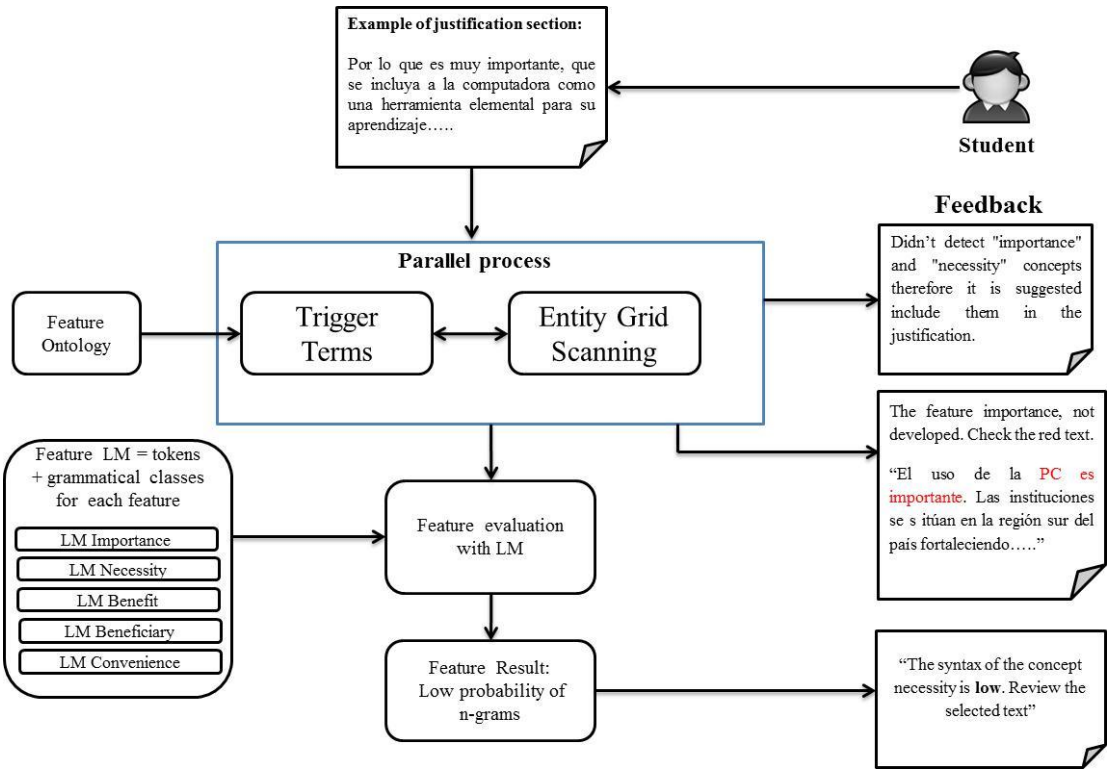

Figure 4. Model to evaluate Justification section of a proposal draft

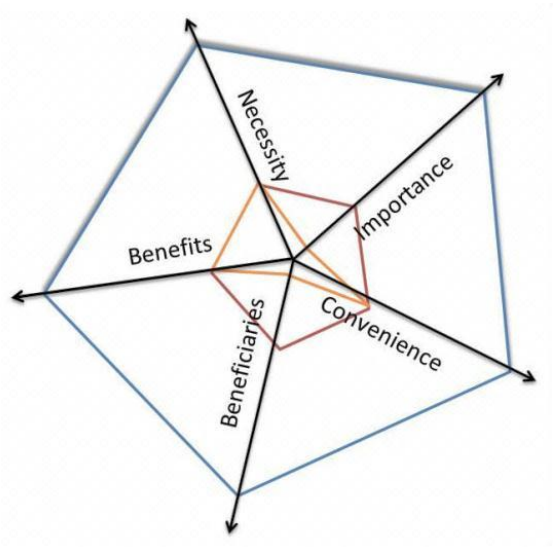

Figure 5. Five Dimensions in Justification section

Five selected dimensions can be perceived with different colors. The blue color represents a proposal draft that complies satisfactorily all dimensions, the color red is a proposal that shows little evidence of the dimensions.

Orange color is a proposal that presents only evidence of some dimensions. Our model

aims to find evidence of more than three dimensions, and also that these dimensions are developed, i.e. not only seeks the presence of terms associated with the dimensions, but also there is an elaborated support of each concept or dimension.

To assess the **Conclusion** section we propose a model that combines the use of language models with Entity Grid technique. We want a model that does not depend only on the corpus.

The technique Entity Grid will allow to identify the entities that are on the objectives and the problem statement of a proposal, and then look for evidence of them in the conclusion section (see Figure 6).

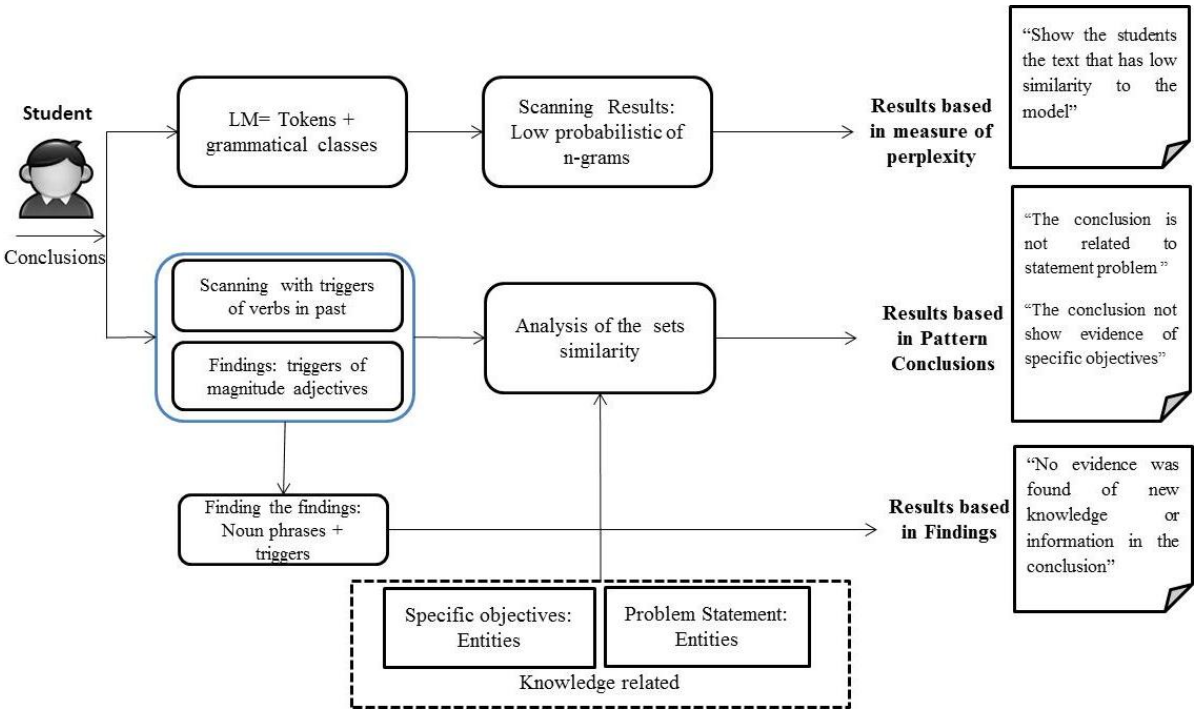

Figure 6. Model to evaluate Conclusion section of a proposal draft

The model is based on the following guide pattern to write conclusions of Teaching and Learning Centre, University of New England. This pattern begins with a reformulation of the problem, followed by key findings, and ending with recommendations (see Figure 7). This guide pattern is similar to a scientific article conclusion, but more extensive.

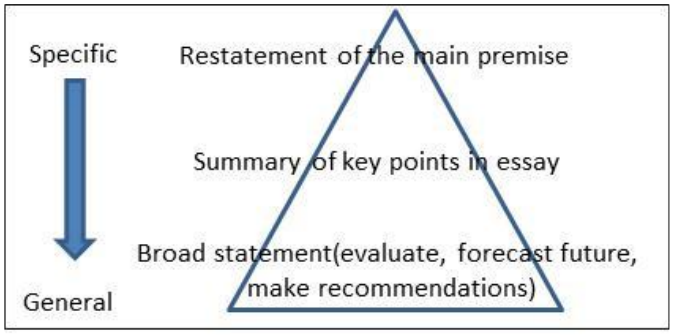

Figure 7. A pattern for conclusions paragraph

Three evaluation schemes has been proposed that combine language models and other techniques such as an Entity Grid to address five sections of a draft. Currently, we are designing other evaluation models for the rest of the sections in a proposal draft.

Furthermore, we are performing the first experiments with language models based on n-grams, for the section of objectives. The results of these experiments will provide us the viability of the proposed models.

11. References

1. Martínez, J., Gutiérrez, D., Hernández F. 2007. Problematic terminal efficiency developed graduate programs in distance mode, in the IPN. *International Congress of Educational Innovation.*
2. Sampieri, R. 2006. *Metodología de la Investigación.* México DF, Mc Graw Hill.
3. Foltz, P., Kintsch, W., Launder, T., 1998. Textual Coherence using Latent Semantic Analysis. *Colorado USA: Discourse Processes*, 285-307
4. Elsner M., and Charniak E. 2008. Coreference-inspired Coherence Modeling. *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*, 41-44.
5. Kuan-Yu Chen and Berlin Chen. 2011. Relevance Language Modeling For Speech Recognition. *International Conference on Acoustics, Speech, and Signal Processing.* ICASSP'11, 5568-5571.
6. Montes y Gómez, M., Villaseñor, L., López, A. 2008. Mexican Experience in Spanish Question Answering. *Computación y Sistemas*, Journal 12(1): 40-64.
7. Pazos, R., Gelbukh, A., González, J., Alarcón, E., Mendoza, A., Domínguez, P. 2002. Spanish Natural Language Interface for a Relational Database Querying System. *In 5th International Conference on Text, Speech and Dialogue.* 123-130. Springer-Verlag London.
8. Oramas, J., De Raedt, L. 2010. Answering Complex Questions in Natural Language using Probabilistic Logic Programming and the Web. *Online Proceedings 22nd Benelux conference on artificial intelligence*.
9. Lapata, M., Barzilay, R. 2005. Automatic Evaluation of Text Coherence: Models and Representation. *In Proceedings of International Joint Conference on Artificial Intelligence*, 1085-1090.
10. Vilarnovo, A. 1990. Text Coherence: Internal Coherence or External Coherence? *ELUA Journal*, 229-239.
11. Foltz, P., Kintsch, W., Launder, T. 1998. Textual Coherence using Latent Semantic Analysis. *Colorado USA: Discourse Processes*, 285-307.
12. Hernández, S. and Ferreira, A. A. 2010. Evaluación automática de Coherencia textual en noticias policiales utilizando Análisis Semántico Latente. *Revista de Lingüística Teórica y Aplicada*, 48(2): pp. 115-139.

13. Kulkarni, S., Caragea, D. 2009. Computation of the Semantic Relatedness between Words using Concept Clouds. *In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 183-188.
14. Barzilay, R., Lapata, M. 2005. Modeling Local Coherence: An Entity-Based Approach. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 141-148.
15. Burstein, J., Tetreault, J., Andreyev, S. 2010. Using in Student Essays. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10).* Association for Computational Linguistics, Stroudsburg, PA, USA, 681-684.
16. *S*elvam, P., Natarajan, A., Thangarajan, R. 2008. Lexicalized and Statistical Parsing of Natural Language Text in Tamil using Hybrid Language Models. *WSEAS Transactions on Signal Processing*, 8(7),1362-1374.
17. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan J. 2000. Language Models for Financial News Recommendation, *In Proceedings of the ninth international conference on Information and knowledge management*, 389-396.
18. Wu, J., Chang, Y., Liou, H., and Chang, J. 2006. Computational Analysis of Move Structures in Academic Abstracts. *In Proceedings of the COLING/ACL on Interactive presentation sessions (COLING-ACL '06).* Association for Computational Linguistics, Stroudsburg, PA, USA, 41-44.
19. Li, Y., Gorman, S. and Elhadad, N. 2010. Section Classification in Clinical Notes using Supervised Hidden Markov Model. *In Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10)*, Tiffany Veinot (Ed.). ACM, New York, NY, USA, 744-750.
20. Nitin Madnan, N.,Heilman, M.,Tetreault, J. and Chodorow, M. 2012. Identifying High-Level Organizational Elements in Argumentative Discourse, *in Proceedings of the North American Chapter of the Association of Computational Linguistics*, June 3-8, Montreal Canadá. HLT-NAACL, 20-28.

12. Appendix A

Key elements in the structuring of research drafts

Documentation of a research project is important, as this is intended to transmit knowledge to others. Good writing is likely to benefit the reader, guiding them on how could apply methods, techniques, theories or simply to clarify any topic related to the study area.

Currently the process of generating knowledge is addressed in many books on research methodology and this has allowed people who are dedicated to developing research, follow a series of steps in a "standardized" way.

This has led to standardization documents are developed that allow in principle to structure a research project and then submit a report of the results. In this case, it seeks to define the key elements that should have a research project, which aims to give support writing to develop.

Reviewing the topic, we found that higher educational institutions and research centers have created their own forms to document a research draft.

This summary presents the research process according to different authors, and some structures to draft, generated by some research institutions. It should be mentioned that book authors of methodology, are used as a reference in research courses in Educational Institutions of Mexico. This will lead to define the key elements that should be included in the structuring of a research draft.

(Hernández, 2006) classifies processes into three research approaches quantitative, qualitative and mixed. The first approach includes the following elements:

Quantitative Approach

1. Problem statement
2. Objectives
3. Research questions
4. Justification
5. Elaboration of state of the art: literature review and construction of a theoretical perspective
6. Hypothesis
7. Define the type of design
8. Data collect

9. Data analysis
10. Results report

Qualitative Approach

1. Problem statement
2. Objectives
3. Research questions
4. Justification
5. Defining the role that the literature reviewed will play in the project.
6. Data collect
7. Data qualitative analysis
8. Qualitative results report

Mixed research combines qualitative and quantitative approaches.

[1] presents the following which should contain a research draft:

1. Problem statement
2. Conceptual framework
3. Research needs
4. Research objectives
5. Hypothesis
6. Dependent and Independent Variables
7. Research design (exploratory, descriptive, causal)
8. Data Collect
9. Measuring instrument design
10. Statistical analysis
11. Conclusions

Elements described above correspond to the quantitative approach.

[2] defines two types of research the quantitative and qualitative. The structure of the quantitative approach includes:

1. Theory
2. Hypothesis
3. Data Production

4. Data Analysis
5. Results

The following comparative table shows the difference between qualitative and quantitative approaches.

Research	Quantitative research	Qualitative research
Relationship theory-research	Structured	Open
Function of Literature	Fundamental to the hypothesis	Auxiliary
Concepts	Operating	Orientation, open
Environment	Manipulator	Naturalistic
Interaction between the researcher and the object of study	Scientific observation, distant	Empathic identification with the object of study
Data collect		
Research design	Structured, closed	Open, is constructed in the course of the investigation
Research instrument	Uniform for all subjects	Varies depending on interest of study objects
Nature of data	Objectives and standardized	Subjective and flexible
Data Analysis		
Analysis aim	Variable analysis, impersonal	Analysis subjects
Mathematical and statistical techniques used	Maximum	Neither
Results		
Data presentation	Relational approach	Narrative approach
Generalizations	Correlations , causal models	Classifications and typologies
Reach of results	Generalizations are sought	Specific

Table1: Table comparative between qualitative and quantitative approaches.

Besides the authors mentioned in the literature review found a guide for the development of research projects, based on the technical standard of the Official Journal No. 313. This guide is used by the University of Guadalajara in the University Center for Health Sciences. Under Article 11 the research draft must contain at least the following elements:

1. Title.
2. Theoretical Framework:
    - Problem statement
    - Background
    - Justification
    - Hypothesis (in appropriate cases)
    - General objective
    - Specifics objectives
3. Methodological design:
    - Materials and Method
    - Study design
    - Sample size
    - Variables
    - Inclusion criteria, exclusion and removal
    - Methods of data collection and statistical analysis
    - Study site
    - Bibliographic references

The University of Ciudad Juarez has a guide to the elements expected of a research project (Lozada, 2005). This document was prepared based on different authors recognized in the area of research methodology such as Corina Schmelkes and Hernández Sampieri. The elements are as follows:

- Title
- Introduction
- Problem statement
- Justification
- Objective

- Background
- Research Questions
- Hypothesis
- Methodology
- Schedule
- Resources
- Basic Definitions and glossary of terms
- Citations and Bibliography

As a result of the review of different materials, elements have been selected coincident between different authors and guides. Also these are the most common elements in the educational institutions of Mexico, where students develop a research project to graduate. Also the structure has been limited to quantitative research, discarding the elements of a qualitative research, because for our study are to review the contents of the computing domain and are mostly used a quantitative approach. Furthermore, in the selection of the eight elements, we have considered the time to develop our thesis.

Below are the basic elements for structuring a quantitative research draft:

1. Title
1. Problem statement
2. Objective
3. Research questions
4. Research Hypothesis
5. Justification
6. Methodology
7. Conclusions

The eighth element “conclusions”, was included instead of "preliminary results" because were not found digital documents that include preliminary results section.

The following defines each of the selected items:

Title: It should be clear and precise, and must comply with at least the following points (Lozada, 2005):

- Avoid superfluous expressions

- Use key words or concepts that relate to the project objective.
- The title should not generate false expectations of the project content
- Analyze the distribution of words in the title, because the order involves perhaps the weight of using a technique
- Write the title in two lines at most
- Do not exceed 15 words

Problem statement: For some authors, the problem statement is to develop and formally structure research idea

The problem statement is caused from a need to make decisions and set the direction of the study to achieve certain objectives, so that relevant data are collected, considering these goals in order to give the appropriate meaning.

Some points to consider are:

- Problem statement should be written in a clear way, accurate and accessible
- It should express the relationship between two or more concepts or variables
- It must involve the possibility of an empirical test

Research objective: They are study guides and are present during the whole development of the research. Their purpose is to indicate to what is intended in the investigation and should be expressed clearly (González and Maytorena, 2004).
The objectives are usually written as grammatical proposition containing:

- The verb must accurately describe an action and is commonly formulated in infinitive.
- The complement indicates the context in which the action is developed.

To compose the target can be answer the questions: What? , How? What for? and his answer was always write infinitive: define, assess, evaluate, etcetera.

According to the verb to be used will be the type of study to be done, either qualitative or quantitative.

Research questions: The questions represent the "what" of research and guide to write the answers that are sought. It is recommended that the questions raised are as specific and precise as possible.

With one or more questions, along with a brief explanation, we can set time limits (time) and spatial (location) of the study and outline a tentative profile of the observation units (individuals, households, newspapers, schools, neighborhoods, phenomena, events).

According to Hernández (2006), the research questions must be met:

- Not knowing the answers
- That can be answered with evidence (observable and measurable).
- Clear and
- The knowledge obtained is substantial, i.e. that knowledge contribution to a field of study.

Research Hypothesis: The hypothesis indicates what is trying to prove. Defined as tentative explanations of the phenomenon under investigation must be made by way of propositions. They are provisional answers to the research questions. Quantitative research with scope descriptive, explanatory or correlation, do have hypotheses. Exploratory research has not hypothesis.

The researcher to formulate the hypothesis unknown if it will be true. The definition of a hypothesis involves the relationship between two or more variables and is supported by organized and systematized knowledge (Hernández, 2006).

Justification: Indicates the reason for the investigation giving the explanation therefor. Through justification should demonstrate that the study is necessary and important. The justification may be drawn around the answers to the following questions:

1. Why and how much is appropriate to conduct this research? or For what will be this research?
2. What are the benefits of this work will provide?
3. Who are the beneficiaries and how?
4. What is the utility of study?
5. Could we fill a knowledge gap?

Methodology: These are the steps and procedures used to carry out the investigation, should include step by step explanation of all aspects needed to play or repeat the research. (Lozada, 2005).

The methodology of the projects includes:

- Techniques and procedures to be used
- Type of research
- The study population
- The sample and the statistical procedure of choice or selection criteria
- Collection instruments and the description of the selection of the data
- Description of the validation instruments. (e.g. pilot test)
- Description of the process of data analysis. (statistical analysis)

Finally, the methodology must be written in future, as a proposal for it to be done.

Conclusions

The conclusions give a global answer to the research question and contemplate the following points (Rolón et al).

- Analysis of compliance of each of the goals of the research.
- Acceptance or rejection of the hypothesis.
- The contrast between the foundations and results analyzing each paragraph of foundations and commenting on the results.
- The constraints that obstructed the investigation.

References

- Corbetta, P. Metodología y Técnicas de Investigación Social. Mc Graw Hill,
- Diario, Oficial. Presentaciòn de proyectos de investigacion. Mèxico DF, 1998.
- Lozada, M. Lineamientos para la elaboracion de proeyctos de Investigación. Universidad de Ciudad Juarez, Ciudad Juarez, 2005.
- González L. D., Maytorena M. Guía de Elaboración y Análisis del Protocolo de Investigación. Universidad de Sonora, Hermosillo Sonora, 2004. Madrid España, 2007.
- Namakforoosh. Metodología de la Investigación. Limusa, México DF, 2007.

- Hernández, S. R., Metodología de la Investigación. Mc Graw Hill, México DF, 2006.
- Rolón A. J., Laria M. J., Rodriguez G. A., Vázquez P. J. *Guía para la presentación del informe de investigación científica.* Comité de tesis de Licenciatura. Universidad Autónoma de Tamaulipas.2007.

13. Appendix B

The Corpus

We gathered 450 documents of theses and reviewed and accepted proposals. The first kind of texts includes documents (proposal and theses) of master and doctoral degree (PG). The second kind includes theses of Bachelor (BA) and Advanced College-level Technician degree (TSU). The corpus domain is computing and information technologies.

Corpus		
Section	Graduate	Undergraduate
Problem statement	40	14
Justification	40	18
Research Questions	40	10
Hypothesis	40	20
Objectives	60	20
Methodology	40	14
Conclusions	40	14

Table 2: Corpus

In total, the corpus consists of 300 examples of graduate level, and 110 of undergraduate level. Each of the sections was collected from theses and digital proposals of higher level institutions in Spanish. 95% of graduate level was of Mexican Universities and 5% of other Spanish-speaking countries and 100% of undergraduate level was of Mexican Universities. All documents are published in official repositories of each University with free access.

We extracted the section of interest of each document and this was stored in a text file. Each section has two files, one for graduate level and one for undergraduate level. It is noteworthy that some collected documents were not included into the corpus, since they did not comply with some required sections. Furthermore, the corpus is constantly growing.

Use of corpus in Lexical Experiments

Both levels (graduate and undergraduate) of our corpus were evaluated considering the three dimensions (density, variety and sophistication) in order to make a comparison of lexical richness among them. The results provide a guideline to be used as a graduate corpus of reference and to establish a scale to evaluate new undergraduate level drafts. For

each section, a scale with following levels was established: Low, Medium and High lexical richness. The High level is defined as one standard deviation (Sigma) above average, Low as one standard deviation below average, and Medium, in between.

The average was obtained for each of the dimensions in every sections, for example, for objective section were averaged the results in density, variety and sophistication dimension of the graduate corpus (60 samples).

Use of Corpus in Global Coherence Experiments

For this experiment we used the corpus of the objectives section, which is composed of 60 examples of graduate level and undergraduate level 20 examples. The first step was to send the corpus (both levels) to three human reviewers with experience in reviewing proposal drafts for classified into low, medium or high global coherence.

Currently, we are working with the other sections of a proposal draft, for which the amounts of training examples and testing, will be different because it depends on the classification of human reviewers.