Linguistic Analysis of Undergraduate Research Drafts

By:

Samuel González López

Thesis submitted to obtain the degree of:

DOCTOR OF SCIENCE IN COMPUTER SCIENCE

at

Instituto Nacional de Astrofísica, Óptica y Electrónica

August, 2015
Tonantzintla, Puebla

Advisor:

Dr. Aurelio López López

Abstract

Academic programs and courses in Mexico often conclude with the elaboration of a thesis or a research proposal written by undergraduate and graduate students. In this process, students are advised by professors who spend time with them. Moreover, students' works must satisfy a set of appropriate structural features pertaining to each section of a thesis. However, according to instructors' experience, the theses exhibit a variety of errors ranging from misspellings to content faults. This research proposes a method to linguistically evaluate essential sections of proposal drafts. The goal is to help two kinds of undergraduate students: Bachelor and Advanced College-level Technician degrees. The method benefits students in their initial drafting, and teachers in their early reviewing. A four level assessment is proposed. The first stage focuses on the lexicon used by the student in his/her draft. The second level seeks to recognize and assess the level of coherence. The third step considers language models intended to classify the particular structure of each element of the proposal. And finally, the fourth level focuses on identifying answers to methodological questions such as "What will you do?" and "How are you going to do it?"; these questions are pertinent for the objectives section of a proposal draft. This thesis presents experiments and results in terms of lexical analysis, global and local coherence, language models and methodological questions on proposal drafts submitted by students. The evaluation of the different models is provided, as well as the results in terms of Kappa agreement.

Keywords: Methodological questions, language models, lexical analysis, coherence, weak sentences, conceptual flow, proposal drafts.

Acknowledgements

I would like to express my gratitude to my advisor **Dr. Aurelio López López** for his continuous support, motivation, and immense knowledge. His guidance helped me to writing the thesis and acquires a discipline to do research.

Besides, I would like to thank to my wife **Francisca Cecilia Encinas Orozco** that with her support and love of every day let me conclude with a dream we both shared during four years.

To my son **Samuel Francisco González Encinas**, that he was unaware that his parents were students, for his young age, became a reason and motivation to continue and finish my doctorate.

I would like to thank the rest of my thesis committee: **Dra. Angélica Muñoz Meléndez, Dra. Alicia Morales Reyes, Dr. Manuel Montes y Gómez, Dra. María del Pilar Gómez Gil, and Dr. Steven Bethard** for their comments and encouragement that contributed to the thesis enrichment.

Finally, I would to express my gratefulness to annotators by their support to tag the corpus used in our experiments: **Claudia I. Esquivel López, Indelfonso Rodríguez Espinoza, J. Miguel Garcia Gorrostieta, Mariela Altamirano, N. Angélica Álvarez Torres and René Castro Morales**. In addition, we want to thank to **April Gloria R. Don Juan** by her support to develop a website of the corpus and the tagging tool.

This research was supported by CONACYT, México, through the scholarship 1124002.

Content

Tables

Figures

Chapter I: Introduction

1. Introduction

Knowledge generation is an important feature of developed countries, and knowledge societies are fundamental in the development achieved by these countries. In Mexico, research and development of new knowledge are supported by Research Centers and some Universities, private or public. Even though, graduate students enrollment increased by 36% from 2010 to 2013, ANUIES[1] reported that only 64% of the students obtained their degree successfully.

Most institutions providing undergraduate programs in Mexico, offer students the option to conclude their studies and get their degree with the preparation of a thesis. From 2012 to 2013 the percentage of students who obtained their Bachelor's degree through a thesis was 71.5%, pursuant to ANUIES. Preparing a thesis is not the only option students have to get this degree. For Advanced College-level Technician degree (TSU), the percentage of graduates in the same period was 60% according to ANUIES. At this level, students finish their program with a short thesis.

Factors affecting the rate of graduation are diverse. A survey[2] applied to students of computer-related careers about problems in preparing the thesis revealed that 8% have problems with document structure [1]. In addition, students reported absence of advice, difficulty in defining the problem to be developed, and the preparation of the thesis project, to name a few. This lack of knowledge leads students to write poor documents.

A study about perception of difficulties faced by students writing discussion section of a thesis showed they are uncertain regarding content and organization of this section. For this purpose, in depth interviews to supervisors (qualified academics) and students were conducted. This information was surprising to supervisors, considering the time and feedback that students received from them [2].

The process of developing a thesis begins by outlining a proposal draft or research project, commonly involving the academic advisor and a student. During this process, advisors spend time reviewing draft formulated by students, and provide recommendations. This becomes a cycle, ending with a proposal that complies with features that have been

[1] National Association of Universities and Institutions of Superior Education in México.
http://www.anuies.mx/iinformacion-y-servicios/informacion-estadistica-de-educacion-superior

[2] The survey was applied to five different generations of students in [1]

established in research methodology books and institutional guidelines. Sometimes this cycle slows down and some of the feedback generated by academic advisors is focused on elements structure of the proposal draft, for instance, a thesis should include a hypothesis or an objective. It is important to note that each element or section of a research proposal has its own characteristics and these elements have to be interrelated [3].

Revising a thesis is a complex task, which requires certain knowledge specific to the area of the thesis by the academic advisor. Also, the advisor must have knowledge of how to write a thesis, in addition have knowledge of the language in which the document is written. Computationally, the modeling of academic advisor knowledge represents a great challenge. In this thesis we took advantage of knowledge expressed in the corpus and a first framework to address the computational challenge was defined, although as a basic formulation.

The work presented in this thesis contributes to the state of the art with methods to help in reviewing theses, taking advantage of the knowledge that exists latent in published theses. The features addressed in this thesis are some that an academic advisor considers when reviewing a thesis [1].

The Students' lexicon is an important element to be considered. In the final document, it must be appropriate. Another feature to analyze in research proposal drafts is studied in [4] and [5], describing 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 (noun) in different syntactic positions and how they are distributed between adjacent sentences, while the semantic approach searches for thematic connection between sentences. Nevertheless, coherence is only one element of several that advisors review.

Syntactic structure of each element on a research project is an aspect to be addressed, *i.e.* how 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 when students write their research project. Some studies have used language models to characterize specific language (spoken or written), supported by probabilistic models [6].

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?”. These questions involve answers requiring the student to reflect and structure various terms, *i.e*. responses do not fit specific data.

In addition, responses are interconnected by internal features, showing logical connections between answers to methodological questions, for instance:

- What will you do? *object* to be achieved.
- What for? The main purpose of the *object.*
- How are you going to do it? Activities or instruments to achieve the *object*.

Currently the search for answers to questions has been studied to find dates, places or names of people in [7] and [8]. The question-answering technique has been used for the retrieval of specific information and part of a query expressed in natural language. Moreover, this task has been addressed with a textual entailment approach, where the question is the text to be linked to a set of answers, selecting the answer that best entails with the text.

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

1.1. Problem Statement

Based on the experience as academic advisors, 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 and 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. Lexicon is an aspect that students must comply from the beginning, but it is often not satisfied, hence interests in this work to address 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 review the draft structure and dedicates less time to examine the content. In addition, progress of student writing is slow.

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

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

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

- What will you do? : *To develop an inductive learning algorithm*.
- What is the purpose of doing it? : *To solve binary classification problems from unbalanced data sets*.

In the answers, the result of reflection process by the student in order to translate his/her ideas is observed. Since the "answer is not known beforehand, it would be difficult to identify such answers. Therefore, it requires a new approach. It is worth mentioning that at this level of identification answer to methodological questions, it is seeks to support the student from a structural approach, *i.e*., this research does not attempt to understand the content of the answers, the purpose of this research work is to provide the student with initial feedback to think about his/her answers to these questions.

Some papers on question answering have divided a complex question into simpler questions by the use of connectors that contain the same question, these connectors are identified by a part of speech tagger [9]. This approach does not seem to be enough for our purpose because the question *What will you do*? cannot be divided into other simpler questions. For instance, the justification section into a thesis requires a deeper analysis by the academic advisor. Authors of research methodology suggest that a justification section has to show evidence that some aspects have emerged, such as importance or needs of the problem. They also mention the benefits of work, as well as beneficiaries. The academic advisor has to scrutinize these concepts when going through the justification section. Several of the review processes are performed easily by the academic advisors, because of his background in the process of thesis review. However, this task represents a challenge for a computational approach, and this research attempts to address it.

Other aspect of interest is reviewing coherence, where a text is coherent if all its parts are connected as a whole [27], which 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 it involves the use of pronouns within a paragraph. For an advisor, it may not be difficult to identify that paragraph is still talking about the same subject and therefore is coherent. This phenomenon of anaphoric elements is difficult to solve computationally. Our proposal seeks to analyze coherence and contribute to the overall assessment of a proposal.

These described issues 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 requirements.

A close study examined 21 doctoral thesis of computing area, taking into account the Bunton model. These theses were written in Spanish and the analyzed section was the introduction. The Bunton model proposes ten steps grouped into three movements in English thesis. The first movement “Establishing a territory”, defines the importance of the issue and provide the background information. The second movement “Establishing a niche” seeks to set the research gap. Finally, the third movement “Occupying the niche” mentions the purpose and the research objectives. Researchers identified in 14 theses the first movement at the beginning of the introduction. However, they identified other new moves that suggest an adaptation of the Bunton model [10]. It is possible to infer with the results that the structure of the thesis written in Spanish is different from the thesis written in English. The work developed in this thesis does not evaluate the introduction section. However, the movements described above are found in some sections of a thesis, for instance, in the objective section.

Thesis analysis has also been addressed with the aim of identifying the sections and communicative purposes in Master and Bachelor thesis [11]. The theses disciplines for this experiment are Philosophy and Linguistics. The authors analyzed 20 theses, they identified a different structure between disciplines, but the academic level did not affect the structure of the thesis. This thesis analyzes documents of graduate and undergraduate levels with the expectation that the structure does not change, considering the institutional guidelines. In addition, this thesis aims to analyze the sections internally and assumes that the student knows the structure that a thesis must include.

Other researchers have addressed some tasks that are developed in this work, but the problem statement is different. For example, the relationship between student essays or the identification of demographic attributes using lexical richness measures. Regarding the concept of coherence, some researchers have addressed the concept in other domains, such as news or student essays. We identified studies that seek matching terms in open answers, given a set of gold standard responses. These studies are detailed in Chapter 3 entitled related work.

1.2. Developed Solution

The developed solution consists of an evaluation at four levels, starting at a basic level as the first filter of the proposed draft, reaching later on a level of complex assessment.

Seven elements of a research proposal are considered as basic elements to evaluate: problem statement, justification, objective, research questions, hypothesis, methodology and conclusions. The elements will be treated differently at each level, some will be processed at all four levels and others only in some of them, due to elements own characteristics. Elements to be evaluated in a draft proposal contemplate the lexical evaluation, assessment of coherence as well as the structure of the seven elements of a draft proposal. Furthermore the problem of the responses to the methodological questions that must be answered to write an objective will be addressed.

1.3. Research Questions

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

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

- How to identify answers to methodological questions such as: What will you do?, What is the purpose of doing it?, 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 semantic and syntactic approaches to improve coherence assessment when reviewing 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 elements 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?

1.4. Hypothesis

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

1.5. General Objective

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

1.5.1. Specific Objectives

- Design a methodology to analyze the vocabulary of each element in a proposal draft.
- Design a method to analyze the coherence 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 the generation of a syntactic pattern.

- Define a method to identify answers to methodological questions of a general objective to help identify the existence of answer to questions, such as: What will you do?, What is the purpose of doing it?, How are you going to do it?.
- Experimental validation of results at each level.

1.6. Contributions

Based on the objectives, we contributed with the following:

- A novel framework to evaluate a students' proposal draft. With this framework we support students and teachers involved in the development of a proposal or thesis in Spanish.
- A new methodology to improve student writing in terms of vocabulary, based on variety, lexical density and sophistication.
- A new method that allows to capture the semantic and syntactic aspects of coherence in the domain of computer science and information technologies.
- A new technique to evaluate the conceptual flow in three sections of a thesis: problem statement, justification and conclusions.
- A new method to evaluate weak sentences in conclusion section, *i.e*. sentences that do not fit into a conclusion.
- A new method for mining the conclusions section, according to three measures: speculation, opinion and the linking between the objective and the conclusion section.
- A novel method to identify answers to methodological questions.
- Theses corpus of graduate (Doctoral and Master degree) and undergraduate level (BA and TSU degree).

1.7. Document organization

The thesis is structured as follows: the second chapter shows basic concepts that were useful for this work development, such as Latent Semantic Analysis (LSA), Entity Grid technique, Coverage Model and Textual Entailment. Chapter 3 describes the work related to the research and gives an overview of the state of the art.

Chapter 4 provides the proposed solution and the methodology used to solve the stated problem. This section describes each of the four proposed levels and the description of corpus gathered from graduate and undergraduate levels.

In Chapter 5, experimental guidelines and results are presented, as well as used corpus and the evaluation of the generated models. Details of collected corpus are provided, considering graduate (Doctoral and Master degree) and undergraduate (BA and TSU degree). Also we include the achieved products of each experiment. Finally, Chapter 6 addresses conclusions and future research

Chapter II: Background

2. Background

In this chapter, theoretical concepts supporting this thesis are presented. First, the concept of Latent Semantic Analysis is detailed. It allows the generation of a semantic space with the best features of the collected corpus. It is also helpful to evaluate coherence aspects. Then, the Entity Grid technique is explained. It provides elements to evaluate the relationship between paragraphs in the sections of Justification, Problem Statement and Conclusion. Finally, the Recognition of Textual Entailment (RTE) is described, including the main applications of this concept. RTE concept was used in the highest stages of the proposed solution. Each concept described in this section is applied at some stage of the proposed solution.

2.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA), at first known as latent semantic indexing (LSI) [12], is an automatic indexing and retrieval technique, which was initially designed for improved detection of relevant documents on the basis of search queries. This is a dimensionality reduction technique based on statistical analysis that allows uncovering the implicit (latent) semantics (structure) in a collection of texts. Afterward, Landauer and Dumais developed the LSA technique [13]. They defined the Latent Semantic Analysis as a theory and a method for extracting and representing the contextual meaning of words in use, through statistical computation applied to a large corpus (documents).

2.1.1. Matrix representation

The information representation of the corpus is the first step of the algorithm used by the LSA technique. This representation involves extracting words frequency in each document and showing a matrix of documents by words. It is a matrix where the columns are a list of words and rows represent a list of documents. The intersection between an element in one column and one row represents the term frequency in the document. For instance:

Consider a collection of three documents: D1: ‘yes yes yes’, D2: ‘no no no’, D3: ‘yes maybe yes’ [14]. The term frequency of the three documents is shown in Table 1:

Table 1. Example of Frequency Matrix[14]

	yes	maybe	no
D1	3	0	0
D2	0	0	3
D3	2	1	0

Then, this matrix is processed to compute weights according to *tf-idf* weight, where *tf* represents the absolute frequency of appearance of a term in a document, and *idf* is the inverse frequency of the term in the documents of the collection, *i.e.* the weight of a term in a document increases if this occurs frequently in such document and decreases if it appears in many (most) of the documents.

2.1.2. Singular Value Decomposition

LSA reveals the (latent) meaning of words, discarding the words occasionally used in specific contexts and focusing on what is common in all contexts [15]. This is achieved by the core process in LSA, Singular Value Decomposition (SVD). SVD allows the simplification of the original matrix to a more manageable number. SVD also diminishes noise or irrelevant information in the matrix. The SVD produces three matrices:

- Orthogonal Matrix U is obtained by linear processing the number of columns in the original matrix A. This matrix represents terms as vectors in the space of words.
- Transpose matrix V^T is obtained by permuting the rows with columns, providing an orthogonal arrangement of row elements. Through this transposition, documents are represented as vectors in the space of words.

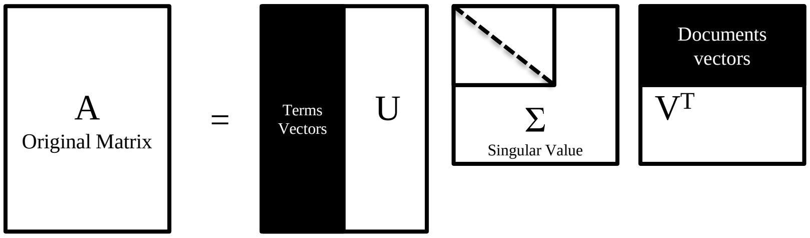


Figure 1. SVD schematic representation [16]

- Diagonal matrix Σ is calculated by linear processing from number of rows, number of columns and number of dimensions in the original matrix *A*. The diagonal matrix represents singular values of *A*. The singular value decomposition of the matrix is illustrated in Figure 1.

Once the three matrices are obtained, a reduced matrix can be generated, but depending on the singular values maintained, it would be a matrix close to matrix *A*, *i.e.* an approximation to A with the most relevant information.

The values of the rows and columns of the reduced matrix are taken as coordinates of points representing the documents and terms on a multidimensional space of k-dimensions, where k represents the original dimensions of the matrix of co-occurrences between words and documents. The number of dimensions (k) is used to calculate the similarities between the text units to compare. Dimensionality is correlated with the occurrence of the terms in the original matrix. [11].

2.1.3. Semantic Space

The semantic space is formed by vectors distributed in an Euclidean vector space, where each vector represents the meaning of words and/or documents produced in the domain of knowledge to which they belong. Such similar vectors represent close latent meanings [17]. For instance, taking as a reference the example in Table 1, the following semantic space is obtained:

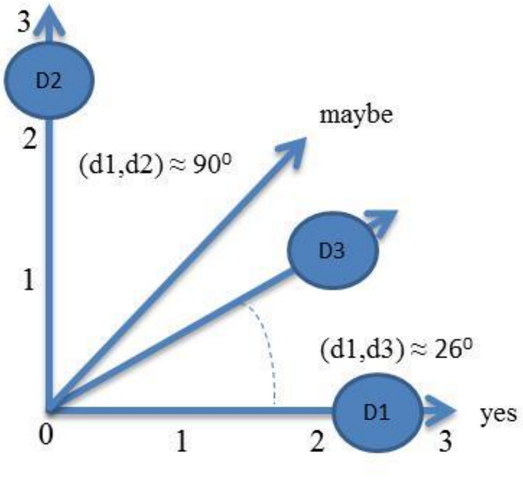


Figure 2. Example of semantic space [14]

Similarity (D1, D2) = D1·D2 = (3,0,0) · (0,0,3) = 0 (orthogonal = 90º)

Similarity (D1, D3) = D1· D3 = (3,0,0) · (2,1,0) = 6

We can observe that D1 and D2 have no words in common and are totally unlike, their vectors are orthogonal, and the algebraic product is zero. However, D1 and D3 share the first dimension (‘yes’) so their vectors are correlated and their inner product is non-zero.
To compute the latent semantic similarity, the cosine of the angle between the vectors is applied to evaluate their closeness in terms of relative frequency or amplitude. The expression for the computation is:

$$\cos\,(A,B) = \frac{A*B}{\| A\| * \| B\|}.$$

where A, and B represent the features vectors. According to this expression, the similarity is 1 when the angle between the two vectors is 0^0, that is, the vectors are pointing in the same direction and are parallel. This result expresses the highest semantic relation in the text. We get 0 when the vectors are orthogonal and correspond to no relation at all.

2.2. Entity Grid

Entity Grid (EGrid) is a technique proposed to represent discourse and then evaluate coherence [18]. A tool based on this technique was used in this work. The technique generates a representation constructed as a two-dimensional array that captures the distribution of entities in discourse across sentences, where rows correspond to the sentences and columns represent the entities of discourse. Cells can have values such as subject (S), object (O), or neither (X). The main idea of this representation is that if the object and subject are present across sentences, the assessed coherence is stronger. For instance, the entity Microsoft appears in sentence 2 as Object and as Subject in sentence 3:

S1: *[The Justice Department]***S** *is conducting an [anti-trust trial]***O** *against [Microsoft Corp.]***X** *with [evidence]***X** *that [the company]***S** *is increasingly attempting to crush [competitors]***O**.
S2: *[Microsoft]***O** *is accused of trying to forcefully buy into [markets]***X** *where [its own products]***S** *are not competitive enough to unseat [established brands]***O**.

S3: *[The case]***S** *revolves around [evidence]***O** *of [Microsoft]***S** *aggressively pressuring [Netscape]***O** *into merging [browser software]***O**.

S4: _[Microsoft]_**S** _claims [its tactics]_**S** _are commonplace and good economically._

S5: _[The government]_**S** _may file [a civil suit]_**O** _ruling that [conspiracy]_**S** _to curb [competition]_**O** _through [collusion]_**X** _is [a violation of the Sherman Act]_**O**.

S6: _[Microsoft]_**S** _continues to show [increased earnings]_**O** _despite [the trial]_**X**.

Note that the identification of the object or subject (Microsoft) roles, is done with a syntactic parser. The EGrid technique generates a model which is built from a specific corpus and this model is used to evaluate new texts. The main idea of this representation is that while the object and subject are present in paragraph being evaluated, the coherence is strong. It is assumed that certain types of subject and object transitions indicate that the discourse has local coherence. Below, there is an example of this technique.

Table 2 shows entities that were extracted from previous sentences. For example, the word *Microsoft* in sentence 1 was labeled as a Subject, in sentence 2 as an Object, in sentences 3 and 4 as a Subject, in sentence 5 the entity was not found, and finally in sentence 6 it was tagged as a Subject.

This transition is shown in the column of *Microsoft* entity reflecting a higher density than the other. According to the authors, there are indicators that the higher the density of the columns is, the greater the coherence level the evaluated text has.

Table 2. Example of Entity-Grid dimensional array[18]

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

In contrast to the semantic aspect, this technique seeks to capture aspects of local coherence, which is something that this work aims to capture measuring the coherence of the proposal drafts.

2.2.1. EGrid probability

It can be observed that the subject with more occurrences in the previous six sentences was the term *Microsoft*. A local transition is defined by the sequence {S, O, X, -} which represents the occurrences of entities and their syntactic roles in adjacent sentences (n). Local transitions can be easily obtained from a grid as continuous subsequences of each column. Each transition will have a certain probability in a given grid. For instance, the probability of the transition [S –] in the grid from Table 2 is 0.08 computed as a ratio of its frequency (six- gray color) divided by the total number of transitions of length two (75).

After calculating all probabilities of the six sentences, the following table would be obtained:

Table 3. Text representation of six sentences[18]

SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	--
0	0	0	0.08	0.01	0	0	0.09	0	0	0	0.03	0.05	0.07	0.03	0.59

To generate the EGrid and conduct experiments to assess coherence, the Brown Coherence Toolkit[3] can be used. The toolkit is a set of C++ libraries and programs for creating and evaluating coherence models.

2.3. Recognizing Textual Entailment (RTE)

There is a connection between two texts if terms in both are associated through some data. In Natural Language Processing (NLP) this relationship has been tackled with the task Recognizing Textual Entailment (RTE).

RTE is defined as a directional relationship between pairs of texts expressions, denoted by T (the entailing "Text") and H (the entitled "Hypothesis").

T entails H if Humans reading T would infer that H is most likely true [19]. For instance:

[3] http://cs.brown.edu/~melsner/manual.html

- *Text1 (**T**):* The SPD party got just 21.5% of the vote in the European Parliament elections, while the conservative opposition parties polled 44.5%.
- *Text2 (**H**):* The SPD is defeated by the opposition parties.

Text1 shows the percentages of voting of two political forces, the Social Democratic Party (SPD) and the opposition parties in Germany. If a person receives this information, it could be inferred that the SPD party has lost the electoral race. Text2 is the result of the inference made by the person. It is noteworthy that inference is a mental process achieved by understanding the read text, besides using prior knowledge. However, the process of automatic inference through computer use is a complex task to perform. Table 4 shows some examples using the Textual Entailment approach, where the fourth column (Judgment) indicates if Text supports the Hypothesis.

Table 4. Examples of Text-Hypothesis pairs[4] [19]

Text	Hypothesis	Task	Judgment
Google and NASA announced a working agreement, Wednesday, that could result in the Internet giant building a complex of up to 1 million square feet on NASA-owned property, adjacent to Moffett Field, near Mountain View	Google may build a campus on NASA property.	SUM	Yes
Drew Walker, NHS Tayside's public health director, said: "It is important to stress that this is not a confirmed case of rabies."	A case of rabies was confirmed.	IR	No
Meanwhile, in an exclusive interview with a TIME journalist, the first one-on-one session given to a Western print publication since his election as president of Iran earlier this year, Ahmadinejad attacked the "threat" to bring the issue of Iran's nuclear activity to the UN Security Council by the US, France, Britain and Germany.	Ahmadinejad is a citizen of Iran	IE	Yes
About two weeks before the trial started, I was in Shapiro's office in Century City.	Shapiro works in Century City	QA	Yes

[4] SUM=Summarization, IR=Information retrieval, IE=Information extraction, QA=Question answering, AA= Answer assessment

Crayfish are territorial and will protect their territory. The shelters give them places to hide from other crayfish. Crayfish prefer the dark and the shelters provide darkness.	So all the crayfish have room to hide and so they do not fight over them	AA	Yes

Currently there are tasks that have been addressed with the approach on textual entailment. Below, there are some of the tasks of interest [19]:

- Question Answering (QA): the task of QA refers to the search for answers to questions such as a specific fact, or information questions (Where?, How? and Why?). The QA problem can be redefined as a problem of Textual Entailment, where the text (T) is the question and the hypothesis (H) is a relational answer pattern (this set of answers is connected with the questions). The selected answer should be considered correct if the corresponding hypothesized answer statement is entailed by the selected passage from which the answer was retrieved.
- Relation Extraction (RE): the goal of this task is to extract text that is connected or that satisfies a proposition. There are two approaches: supervised and unsupervised. In the first approach the training set is annotated with mentions of relation, and their related arguments. For instance:

  ***Microsoft**, the private equity group that was founded by **Bill Gates**….*

  Microsoft and *Bill Gates* are entities connected by a relation "founded". The mentions of the relation help to identify connections between the proposition and the extracted text. In the second approach a template is provided and it specifies the text extracted features and the arguments. In this approach the template corresponds to the Hypothesis.
- Answer Assessment (AA): this task is relevant to this work. The objective is to recognize whether a students' answer to an open question is adjusted to the correct answer or whether the answer contradicts the correct response. Student response and the correct answer are fragmented into short propositions. The entailment connections between these propositions are used to obtain the score of student answer.

2.3.1. Coverage Model: Token Level Similarity

The approach named Token Level Similarity is used to resolve a basic textual entailment case where the hypothesis is expressed directly in the text T. The main idea of this approach is to verify if the hypothesis is part of the text by counting the number of tokens that they have in common. This approach has been used in this work under the name of Coverage Model. Table 5 illustrates an example that seeks to determine if T implies H using the Token Level Similarity:

Table 5. RTE example[19]

T_1 implies **H_1**
T_1: The **four refineries located** in **Gulf** of **Mexico** appear to be the ones hardest hit by the water and wind that accompanied the **hurricane Isaac**.
H_1: **Hurricane Isaac** caused damages to **four refineries located** in **Gulf** of **Mexico.**

It can be observed that the common words are presented in bold in the hypothesis. With the coincidence of terms, it is possible to generate a scale to decide if there is enough evidence to predict that T entails H. An option for building the scale would be counting the matching tokens between T and H, divided by the number of tokens H.

In this approach, only content words are considered, that is empty words (prepositions and determiners) are ignored. For the previous example a value of 7/9 = 0.77 would be reached, this value can be considered as high since tokens coverage is over 50%, *i.e.*, the Text T implies the Hypothesis H.

Under this approach and the example described above, the decision process used by the RTE decision scheme is described below. RTE scheme of decision involves three steps: the first is named Candidate Alignment Generation, the second is the Alignment and the third step is the Classification [19].

- Candidate Alignment Generation: the first step is to identify the tokens or similar phrases among the text *T* and hypothesis *H*. The token-token comparison is performed to determine their similarity. In the previous example, the token "Gulf" was identified as a similar term in both *T* and *H*. However, if the token "Gulf" is replaced by "Bay" in *T*, the similarity with *H* will be null. This case would be resolved using resources of synonyms. The output of this step is a list of anchors, *i.e.* an anchor represents the link between the token of *T* and the token of *H* (with a binary value of 0 or 1).

- Alignment: this step seeks to align the tokens between *T* and *H*, choosing the token of H that best matches in *T*. In this step a list of the best anchors is obtained.
- Classification: to decide if *T* implies H (given a set of best anchors), the classification can be processed as an average similarity of relations between *T* and *H* (edges) and then compared to a threshold.

It is noteworthy to state that the token level similarity model could be more sophisticated adding similar verb structures or phrases. Even the token level similarity model could perform token-token alignment considering syntactic structure of the aligned tokens. This representation's main strength is the fact that the decision function will be modeled in terms of similarities between *H* and *T*.

2.4. Kappa Measure

Kappa Cohen is a measure of agreement between two categorical variables, X and Y [20]. For example, Kappa can be used to compare the ability of different evaluators to classify documents in two or more groups. Kappa is calculated from the observed and expected frequencies on the diagonal of a square contingency table. Below, the Kappa Cohen equation is described.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) is the observed agreement between observers
P(E) is the hypothetical chance of agreement by chance.
K=1 means that evaluators were in complete agreement.

$$P(A) = \frac{number\ of\ agreements}{number\ of\ agreements + number\ of\ disagreements}$$

$$P(E) = \sum_{i=1}^{n} (p_{i1} \text{ x } p_{i2})$$

n=categories
i= number of categories (1..n)
$P_{i1=}$ proportion of occurrence of the category i for observer 1
$P_{i2=}$ proportion of occurrence of the category i for observer 2

The following table shows the interpretation of the results of Kappa:

Table 6. Kappa interpretation[20]

Kappa value	Interpretation of agreement
< 0	Poor
0.0 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Chapter III: Related Work

3. Related work

The stated problem requires the application and development of techniques for Natural Language Processing (NLP). Identifying answers to methodological questions involves challenges in NLP area and also the exploration of existing techniques. Related works that have been developed by researchers are presented in this section. They have addressed different issues that are the closest studies to this work. These studies are grouped under different concepts that characterize a student text such as Lexical Richness, Coherence, Syntax and Inference.

3.1. Lexical Richness

Richness in vocabulary (Lexical Richness) is an academic competence that students acquire while advancing in their education. Therefore, when the students reach the college level, their writing skills, and particularly their lexical competence, should be above those of an elementary and middle school student. This richness is associated with the students' ability to understand different concepts, allowing students to structure sentences that are rich in vocabulary and meaning. The concept of lexical richness has been addressed by various researchers, especially in the evaluation of English essays. Also, lexical richness has been used to define the quality of the document, *i.e.* the ability of a writer to use vocabulary properly [21].

Colleges and universities in Canada take into account the results obtained by students in proficiency exams of different areas. One of them refers to the domain of English (*e.g.* English Language Arts 30-1/2 in Canada) while other refer to the domain of mathematics (for instance PMAT 30) [22]. This study conducted at the University of Calgary for Non-Native English Speaking (NNES) students, aimed to relate the academic success of students with lexical richness. One of their main objectives was to compare the lexical richness of NS (Native English Speaking) and NNES (Non Native English Speaking) students with their academic performance. The authors used parameters such as the English language proficiency and academic achievement in higher education, vocabulary use and academic writing, faculty perceptions of NNES writing, and also lexical measures such as vocabulary size, knowledge of words, and corpus linguistics; recurring later on to evaluate common measures of lexical variety and doing vocabulary profiling.

Douglas [22] found that NS students on the English-30 test have higher scores compared to NNES students, but in the PMAT-30 results are significantly higher for NNES students. Another result is that students with higher lexical richness measures tended to perform better on the Effective Writing Test (EWT) while students with lower lexical richness measures performed poorly on the EWT. Finally, the author concluded that results suggest that students with appropriate vocabulary (varied and accurate), have excelled in their studies; while students with a general vocabulary, (repetitive and an uncontrolled set of vocabulary) showed inferior academic performances. This conclusion supports the efforts aimed to improve the vocabulary of students in their research drafts. Besides helping the academic advisor on focusing on the content of drafts, the skill seems to have a collateral beneficial effect.

The relationship between lexical richness and the quality of essays was studied in advanced English students at a Swedish University [23]. The essays were produced by 37 students of English. The measures used were essay grade (assigned by human reviewers), course grade, and vocabulary knowledge (size of a students' vocabulary). All students wrote four essays in three literary genres (poetry, fiction, and drama), and one essay on a topic of their choice. The students were graded into three levels: Fail (F), Pass (P), and Pass with distinction (PwD). To review the essays, 20 teachers were required. The reviewers evaluation results were called lexical frequency profile (LFP). After the experiments, no relationship was found between lexical richness and quality of the essays. One explanation is that reviewers focused on the content of the essay and grammar features rather than on the lexical features. They suggested using LFP results as a diagnostic tool to identify students with poor writing. Also they concluded that teachers should encourage students to improve their knowledge of vocabulary to write essays in English, as the current efforts of our work aspire.

Several researches have proposed and used different quantitative approaches to measure the writer ability to use vocabulary (lexicon) in compositions; most of them with different goals. For instance, one of them is to measure the sophistication of some papers using text word lists. In [24], the authors used a list of 3000 easy words. For Spanish, some studies used the list provided by the SRA (Spanish Royal Academy) of 1,000, 5,000 and 15,000 most frequent words.

In [25], Roberto et al. used 32 lexical measures to predict demographic attributes, such as age or gender, regardless of the domain. Those measures were grouped in three dimensions: lexical variety, lexical density and sophistication. The authors used the corpus Hopinion with more than 18,000 Spanish texts of opinion from the TripAdvisor site (www.tripadvisor.es). From these texts, only 1,911 items were selected, because they contained morphological and demographic information. The texts were rated on attributes such as gender, age and country, generating different classes within each attribute. For instance, the gender attribute had two classes: male and female. The first experiment was to implement the 32 measures for each attribute and dimension.

With the generated measures, different classifiers were trained using 90% of the texts for training and 10% for validation. In the second experiment, the group of classifiers for each dimension was trained in order to know if any of the dimensions was a better predictor than any other demographic attributes. The first results showed evidence that the classifiers achieved the best prediction rates using the age attribute and then the country attribute. In the second experiment, the sum of dimensions was more effective in predicting using demographic attributes. It was also found that attributes such as: age, gender and users country are poorly related to sophistication. The measures were grouped in density, sophistication and lexical variety, in the same way that the methodology developed in this thesis to evaluate the lexical richness. However, one difference was that a resource of 1,000 most common words was used to compute the measure of sophistication.

Similar to the study previously explained, a discriminant analysis was performed with the goal to identify whether or not the lexical features (word, sentence, lexical overlap, connectives, and lexical diversity[5]) predict with low statistical significance and high essay quality [26]. A discriminant analysis is a statistical multivariate technique (regression analysis) that predicts group membership using a series of predictor variables.

Results of discriminant analysis for low and high quality using only variables of syntactic complexity, lexical diversity and word frequency predicted correctly 52 of 80 essays of the training group, and 28 out of 40 essays from the test group. The model obtained 67% of accuracy.

[5] These features were extracted with Coh-Metrix tool. http://cohmetrix.com/

Authors concluded that it is not enough to say that the variables used in the discriminant analysis helped to differentiate the two groups. However, higher scored essays were more likely to contain associated linguistic features with text difficulty and sophisticated language. This work is looking for lexical analysis to identify frequent deficiencies in student writings, such as excessive use of empty words and certain terms, or deficient knowledge of technical terms.

3.2. Coherence

Many text definitions include coherence as a necessary feature. A formal definition given in the work of Vilarnovo [27], 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.

Coherence is classified based on its scope: Global and Local. Global coherence means that a document is related to a main topic, *i.e*. it is not consistent when its elements have no such main topic (semantic aspect). Local coherence is defined within small textual units [28]. Recently, Skogs [29] reported a study of different factors conducing to cohesion and coherence in texts coming from student discussion forums (syntactic aspect). An exploration on how foreign language learners express cohesion and coherence in their writings was reported in [30], employing topical structure analysis. An analysis of several methods for assessing coherence in the context of automated assessment of learner's responses was given in [31]. In [32], authors defined four aspects related to local and global coherence (Relatedness to prompt, relatedness to thesis, relatedness within segment, and Technical Errors), one of which is connected to the topic developed in the essay about the required topic by the teacher. Despite the focus on local coherence, in [33] Miltsakaki and Kukich highlighted specific areas of research for NLP in essay scoring. None of these studies of coherence discussed proposal writings. They are predominately studies on how to grade essays already written, *i.e*. not to support directly the writing process.

3.2.1. Global Coherence

Several previous works have focused on evaluating educational aspects using the Latent Semantic Analysis (LSA) technique. In the educational field, different kinds of documents are generated, such as documents written by teachers related to learning activities, student essays or textbooks [34]. This work focuses on proposal drafts of undergraduate students, specifically in the Spanish language.

In the study by Foltz, et al. [4], they evaluated the textual coherence using LSA technique. This work shows the coherence prediction by analyzing a set of texts (statement by statement) of four texts, with a 300-dimensional semantic space, which was constructed based on the first 2,000 characters of each of the 30,473 articles of the Encyclopedia of American Academic Groliers. After separating 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 [4] is whether or not the LSA technique is a model of text level knowledge of an expert or novice. The authors suggest that it depends on the training the technique has received in the application domain. This technique focuses on the latent semantic aspect, which would be a relevant aspect to this research.

In the study by Ferreira and Kotz [35], the authors evaluated 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 also used the technique of Latent Semantic Analysis. First compiling a corpus in the police news domain which served to train the system, and from that collection, the system measured the coherence of the news.

The expected result was that the coherence system will be close to the evaluation done by a journalist and a Spanish teacher. The results of six texts of evidence showing in Table 7:

Table 7. Results of coherence level in six police news [35]

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 6 shows that the values of level of coherence between the machine and the journalist are close, but between the machine and the Spanish teacher there is a considerable difference. Text 3 shows a clear difference in the level of values of coherence.

The authors explained that was because in the generated corpus 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 text 3. Finally, the authors concluded that the results were positive, since in the newsroom there was 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 [36], the authors intended 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 [36] is that it is not restricted to a particular field of knowledge. A tool that is available on the web[6], 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.

3.2.2. Local Coherence

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 [37].

In [38] the researchers conducted an evaluation of different techniques used to measure coherence in text, considering semantic and syntactic approaches. Techniques that evaluate semantics are based on words, and distributional similarity measures of WordNet, such as, HStO, Lesk, JCon, Lin, and Resnik. For the syntactic approach the authors selected Entity Grid.

In this experiment [38] it was considered the human judgments as the highest level and that was the basis for comparison with each technique. Also, a multiple linear regression to determine the degree of correlation existing between each technique was applied.

Table 8 shows the correlation results, measured with the Pearson coefficient that each of the models reached, with respect to the results of human judgments. Latent Semantic Analysis reached a correlation of 0.230, Entity Grid 0.2446 and the HStO 0.322 being this

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

one the highest correlation obtained. 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.

Table 8. Correlation between human judgment and coherence techniques[38]

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

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 students' essays on three different populations [39].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.

A related study with local coherence of students learning English is presented in [40]. Authors argued that essays paragraphs have an internal flow, *i.e.* each paragraph connects with the adjacent paragraph. If paragraphs are connected in ascending order, the essay parts are coherent between them (see Figure 3).

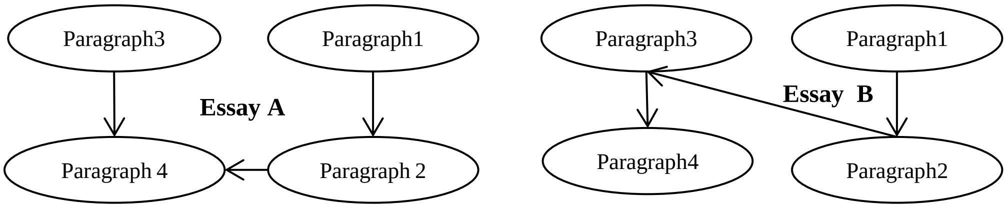


Figure 3. Essay A (left) with a low grade and Essay B (right) with a high grade of semantic flow [40].

O'Rourke, S., and Calvo [40] used singular value decomposition, where each paragraph of the essay was represented in a vector space, and then they measured the distance between vectors to determine paragraphs semantic proximity, that depended on their topics. The authors also implemented the Non-negative matrix factorization method, and found that this was suitable for topic flow analysis. Despite the fact that topic flow in essays was small, it was present. They concluded that is possible to obtain better semantic flow on collections of essays with more significant quality differences.

In this thesis, for example, in the Problem Statement section, the set of sentences that integrate each paragraph are interconnected by the same central topic. This flow of connections provides an adequate sense of what the student seeks to address in the research proposal and reflect the connection of all parts as a whole.

One of the lines to work in this thesis is the fusion of some of these techniques, in order to assessment the Global and Local Coherence in early versions of proposal drafts.

3.3. Syntax

Another component of interest is the syntactic characterization of each element of a proposal draft, through the use of language models. In the study by Selvan et al. [41], a lexicalized and statistical parser is presented for word processing of a regional language in India (Tamil). The authors used language models to generate probabilities associated with each word.

The researchers 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, the authors used n-grams, where the probability of each word depends on the n-1 word, the best results was reached with n = 3, then was combined the language model, *i.e.* the statistical approach with the structural approach, the latter refers to the grammatical structure that a sentence has. Figure 4 shows the parser design.

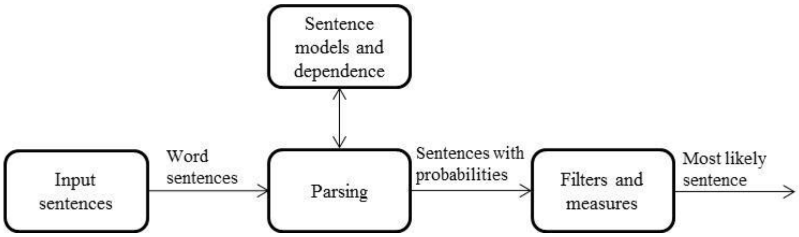


Figure 4. Lexical and statistical parser[41]

Similarly to [41] this research seeks to identify the syntactic patterns in the elements of a proposal draft, through the use of language models. The following works propose the use of language models to generate financial recommendations, specifically looking for financial news stories that might influence the behavior of markets. The work proposed a scheme in which two types of information were retrieved to generate the language model. First, information about product prices is retrieved, from which the price trends are built. In parallel, the authors collected items related to finances, and used a collection of 38,469 articles from Biz Yahoo [42]. This thesis research attempts to 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.

In the study by Wu et al. [43] structured movements on articles abstracts were analyzed. First the authors 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 served as seeds to train a Markov model. Then, the authors automatically extracted collocations in order to find phrases that represented rhetorical moves. For example, the collocation “paper address” was found in the training corpus and was labeled with the type of movement “P [7]”. With this collocation were possible tag new sentences.

Table 9. Example of found collocations[43]

Collocation	Move type	Number of collocation 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

Another procedure performed in this work [43] was to expand the collocations in order to capture similar movements, although the collocation could appear separately. 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 as “P” movement. Table 9 shows some examples of collocations found by authors.

[7] Background (B), Paper address (P), Method (M), Result (R), and Conclusion(C)

The training corpus contained 20,306 abstracts with 95,960 sentences obtained from the 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 this work [43], but it is necessary to consider that there is not a large corpus which enables finding valid collocations. However, one can consider also the way the authors used the Markov model.

Other studies have addressed the identification of sections within documents. One of the study conducted by Li et al. [44] was focused on the classification of sections within clinical notes, implementing a HMM (Hidden Markov model). In this work [44] the researchers used a corpus of clinical notes of New York-Presbyterian Hospital with 9,679 notes and 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, an HMM with 15 states was trained, 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% notes without labels section.

Also, the authors built a dictionary of labels, for example words “Treatment plan”, “impression/plan”, and “assessment and plan” were mapped to “A/P”. To evaluate the accuracy of the dictionary, the researchers used 120 clinical notes tags from two physicians, reaching 97.36% accuracy. Finally, the corpus was divided in 78% for training and 22% for testing, to evaluate the model-classifier. The results of F-Measure were about 90%, statistically above the baseline, which was about 70%. For each note, the accuracy reached 70% in HMM compared to 19% for the baseline.

Within these results, the authors found that section STUDY qualified for error in LABS section by 10.24%, this was because they are adjacent. Also CC section was classified in

A/P by 23.36%. It can be said that the method is sensitive to adjacent sections. This thesis could help in one part of this research thesis, but emphasizing that the purpose is not to identify the sections of a research proposal, but to evaluate the structure of each section.

3.4. Inference

The inference process attempts to reach a conjecture from current information. The Spanish Royal Academy defines the term infer as *deduce something and get a consequence*. An inference is defined as any assertion that the reader comes to believe is true, as a result of reading the text, but not established by the reader previously, and that is not explicitly stated in the text [45]. Moreover, the inference is generally perceived as the process by which new consequences conclude from the given information [19].

The inference process is inherent when writing a thesis. It is mainly used by students during the analysis of the results of their research, allowing improvements or making changes to the implemented methods. Inference is also required in early stages of drafting the proposal, since after understanding the problem to solve and defining the research questions, the student states the general objective that could solve the problem statement.

In addition, the review of techniques related to the problem helps students develop their methodology. Also the knowledge learned prior to specific objectives development will allow the student to solve the problem statement.

Studies under the concept of inference through Recognizing Textual Entailment (RTE) task [19], try to solve problems involving the identification of truths from fragments of texts, *i.e*. from a text T1 it can be inferred a second text H1. H1 entails T1.

The Answer Assessment (AA) tries to evaluate student responses to open questions. The AA task is close to the problem of identifying answers to methodological questions that this thesis seeks to solve. Identifying answers to questions managed by an intelligent tutor has been approached from the perspective of machine learning [46]. The student answers questions and then these are evaluated to see if he/she understood the concept.

As detailed in Table 10, the student gives an answer describing the functionality of a body part. The text 1 corresponds to R and A corresponds to Hypothesis. The authors generated a corpus of annotated answers finely, specifically the where and how the response did not meet expectations. In addition, each reference response was decomposed into a dependency tree. The authors used a machine learning approach.

Table 10. Question (Q) with reference answer (R) and student answer (A)[46]

Q: Dancers need to be able to point their feet. The tibialis is the major muscle on the front of the leg and the gastrocnemius is the major muscle on the back of the leg. Describe how the muscles in the front and back of the leg work together to make the dancer's foot point.
R: The muscle in the back of the leg (the gastrocnemius) contracts and the muscle in the front of the leg (the tibialis) relaxes to make the foot point.
A: The back muscle and the front muscle stretch to help each other pull up the foot.

The results obtained from this approach had 75% of accuracy, which is a percentage close to the gold standard. This work is close to the purpose of the present thesis since it deals with identifying answers to methodological questions. It is also similar because there is an interest in seeking an answer that reflects that the student has understood a question.

The problem of the evaluation of short answers as an inference problem is addressed in recent studies [47]. Horbach et al. analyzed whether or not the student has understood a text through the evaluation of the response the student provides; in this case, it seeks to validate the contents of the answer. In this study [47] besides aligning the answer to the original text, it attempts to identify which part of the text is responsive to the question under discussion. The work focuses on students learning a foreign language (Germany). The grammar assessment is omitted.

Table 11 depicted a paragraph of the story from Germany. One can notice on the table that the correct answer fits with one sentence of the full text (in gray). The level of agreement between annotators to tag the student answers with correct labels was 70% and 65% for incorrect answers.

To entail the student answer to the original text, the first approach used by the authors was the answer-based model. This model begins with the alignment of the terms of the students' answer with the target answer, question or reading text using tokens, chunks and dependency triples. From the alignment process 15 features were extracted to introduce them in a classifier (for example the keyword overlap, target token overlap, learner token overlap, lemma match, synonym match, type match and target triple overlap). The proposed hypothesis was that if the students' response was connected in some proportion to the original text the answer would be correct.

Table 11. Example with questions and answers [47]

Text: Schloss Pillnitz: This palace, which lies in the east of Dresden, is to me the most beautiful palace in the Dresden area. One special attraction in the park is the camellia tree. In 1992, the camellia, which is more than 230 years old and 8.90 meters tall, got a new, moveable home, in which temperature, ventilation, humidity, and shade are controlled by a climate regulation computer. In the warm seasons, the house is rolled away from the tree. During the Blossom Time, from the middle of February until April, the camellia has tens of thousands of crimson red blossoms. Every year, a limited number of shoots from the Pillnitz camellia are sold during the Blossom Time, making it an especially worthwhile time to visit
Question: A friend of yours would like to see the historic camellia tree. When should he go to Pillnitz, and why exactly at this time?
Target Answers: • From the middle of February until April is the Blossom Time. • In spring the camellia has tens of thousands of crimson red blossoms
Learner Answers: • [**correct**] He should go from the middle of February until April, because then the historic camellia has tens of thousands of crimson red blossoms. • [incorrect] Every year, a limited number of Pillnitz camellia are sold during the Blossom Time. • [incorrect] All year round against temperature and humidity are controlled by a climate regulation computer.

The second approach in [47] was the Text-based model, which considers the connection between student answer and target answer, unlike the previous approach which considered the original text. In this approach, only an intersection between sentences is evaluated (student and target answer), so the answer will be correct if it gets a high value of intersection between tokens. Also the researchers conducted experiments combining both approaches. Finally, the simple text-based model improved the classification accuracy over answer-based model.

The two works described above seek to assess answers to questions for evidence that the student has understood a concept or text (for learning a second language), this scheme is similar to the methodological questions raised in this work. It can be seen that the answer contains information that can be linked by the terms of the original text or questions.

In contrast, identifying answers to methodological questions (which this thesis aims to do) is not just a matter of the similarity of terms the student writes in an objective. The answer to the question “What will you do?” does not share similar terms with the question.

However, the techniques identified under the concept of inference will support the development of a solution to the problem statement in this thesis.

Chapter IV: Developed Solution

4. Developed Solution

The strategy to solve the stated problem is based on a four-level analysis using natural languages processing techniques. The idea is to process students' proposal drafts, starting from the basic level (*e.g.* lexical analysis) and afterwards working with the complex level (*e.g.* finding answers to methodological questions). At the bottom of Figure 5, one can see the seven elements of a research proposal draft considered as basic elements for evaluation. Gray bars represent the level where each element can be subject for analysis.

The objective section is required element that can be analyzed at the fourth level; given their formulation, this element allows answering certain methodological question. At the first level, lexical analysis is conducted in order to know if the student is using diverse vocabulary and an adequate balance of content words.

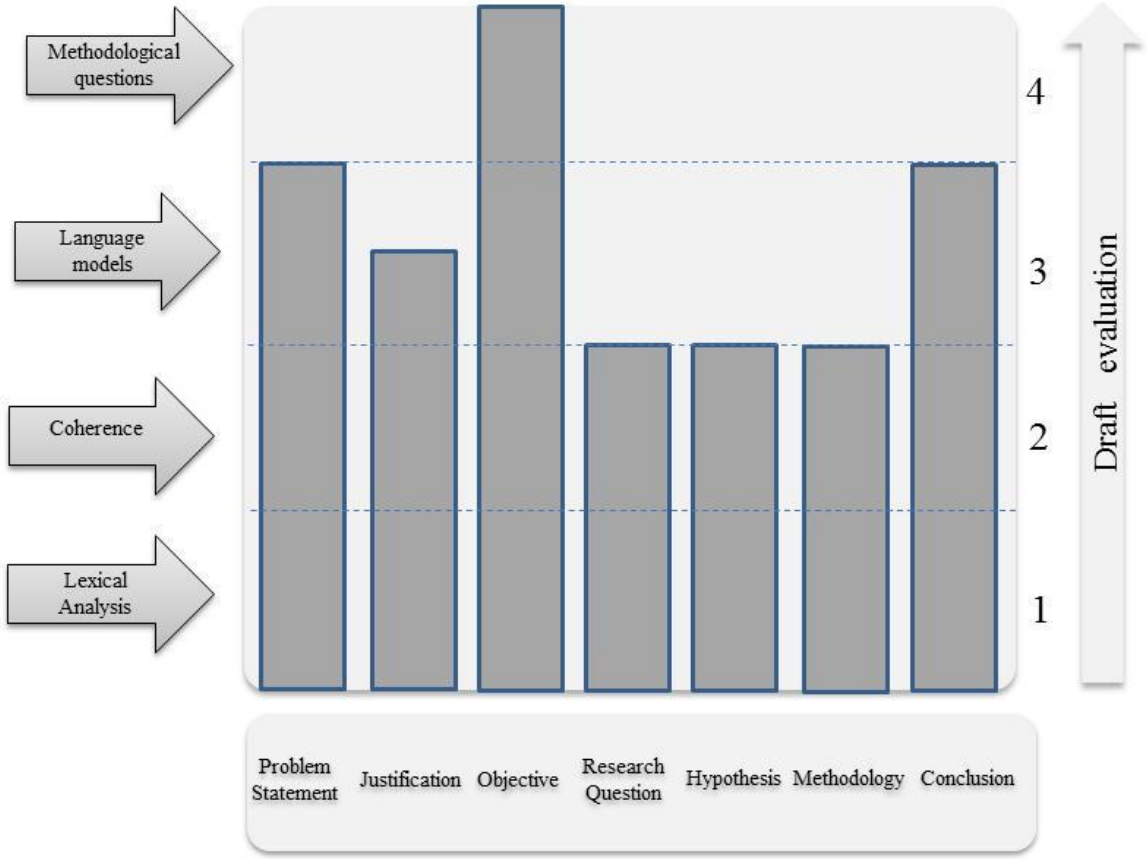


Figure 5. Four- level evaluation

It is also aimed to determine the vocabulary richness based on three dimensions: lexical variety, 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 the evaluation begins; it provides a perspective of basic language level.

The second level focuses on evaluating global coherence and the flow of concepts (implying local coherence) and it is based on the combination of semantic and syntactic approaches. This level attempts to capture if elements of the proposed research are semantically coherent to the area of computing and information technologies, but it also revises that elements themselves are connected, by incorporating the syntactic approach. This combines both coherence approaches considered in previous studies, which showed that the used techniques capture complementary aspects of coherence [37].

The third level addressed the task of evaluating a proposal draft according to language models capturing the syntactic structure proper to each element. This level looks for thematic independence, *i.e.*, we want to register what kinds of discourse elements are being used and how they are used, such as the use of verbs, adverbs, nouns. This level is above the thematic aspect and emphasizes in the syntactic elements characterizing 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, since previous work in question answering is intended for a specific level and factual data. This is the highest level because it does not only look for syntactic or lexical features, but it also intends to identify answers to questions that involve a process of reflection from the student, as illustrated above.

The four levels have an independent performance, thus each student could use each level according to their need. However, the developed four-level framework allows that the thesis assessed at the first level can be the input of the second level.

It is noteworthy that one level cannot achieve the solution to the problem statement in this thesis. For instance, the global coherence evaluates a latent semantic aspect, while local coherence assesses the internal connection of the document. These two approaches are complementary, so the student document will be refined at the following levels.

Therefore, the proposal drafts written by students 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.

4.1. Methodology

An assessment at four levels was proposed to attack the problem statement of this thesis. The first level evaluates lexicon, the second focuses on conceptual flow and global coherence; the third addresses the language models that characterize the proposal elements and the highest analysis concentrates in identifying answers to methodological questions. Each level demands specific techniques, given the varied nature of features to study. The following points describe the steps taken to solve the problem statement.

4.1.1. Proposal draft elements

To describe the general problem, it was necessary to select the proposal draft elements considering the computational viability. For this purpose, several books [1] [3] on research methodology were reviewed as well as institutional guides by universities. The sections selected were: Problem Statement, Justification, Research Questions, Hypothesis, Objectives, Methodology and Conclusions [3]. Two elements not covered in this thesis are the title and the results section, for instance.

4.1.2. Corpus

Gather a corpus including research project proposals and theses to facilitate the identification of features of interest. Proposals for research projects are gathered considering that these are written in Spanish. The corpus will also be helpful to carry out experiments. In total 468 theses were collected. The corpus is detailed in section 5.1.

4.1.3. Lexical Methodology

For each of the elements on a proposal draft was designed and implemented a lexical methodology. The methodology developed is detailed below:

To evaluate the seven elements contained in a research proposal, a computational model containing three lexical dimensions was proposed. The first step in the model considers preprocessing of each element. Each section in this module is processed with the Freeling[8]

[8] http://nlp.lsi.upc.edu/freeling/

tool to obtain the word stems, converting the analyzed word in its singular form, grouping similar terms, and allowing a fast lexi**c**al analysis (see Figure 6).

Another step in the preprocessing of the text was filtering and removing stop words from a list of 209 words provided by the Natural Language Toolkit. Stop words include prepositions, conjunctions, articles, and pronouns. After this step, only content words remained, which allowed calculation in three dimensions. Punctuation marks were omitted. In the section evaluation module, three methods for dimensions calculation were included. The first procedure is the lexical variety which seeks to measure student ability to write their ideas with a varied vocabulary (see Table 12).

Table 12. Measures to compute lexical richness

Dimension descriptions		
Dimension	Label	Computed as
Variety	LV	*Tlex/Nlex*
Density	LD	*Tlex/N*
Sophistication	LS	*NSlex/Nlex*
Tlex: Unique lexical terms *Nlex*: Total lexical terms *NSlex*: Tokens out of a list of common words (SRA) *N*: Total tokens		

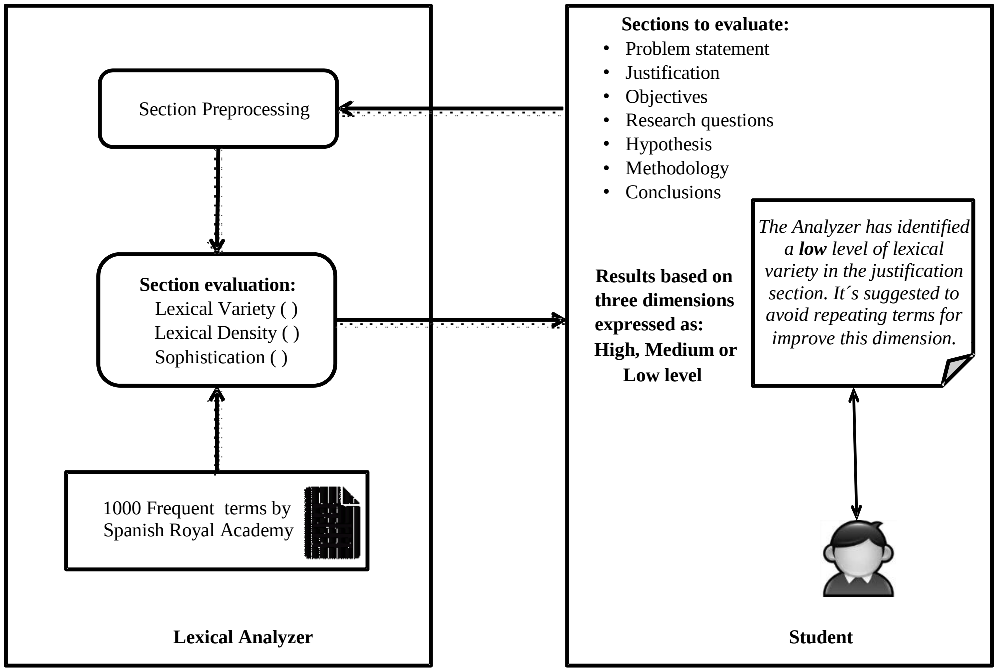

Sections to evaluate:
• Problem statement
• Justification
• Objectives
• Research questions
• Hypothesis
• Methodology
• Conclusions
Section Preprocessing
Section evaluation:
Lexical Variety ( )
Lexical Density ( )
Sophistication ( )
Results based on three dimensions expressed as: High, Medium or Low level
The Analyzer has identified a low level of lexical variety in the justification section. It´s suggested to avoid repeating terms for improve this dimension.
1000 Frequent terms by Spanish Royal Academy
Lexical Analyzer
Student

Figure 6. Lexical Richness Evaluation Methodology

The second module refers to computing lexical density, whose goal is to reflect the proportion of content words with respect to all words that were employed, *i.e*. if the text has a good level of content. Finally, the sophistication method attempts to reveal 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 this case, 1000 common words, according to SRA).

A related work has identified that the dimensions selected in this experiments achieved a good performance, for instance in [57], a study was presented, comparing two of the measures most used to evaluate texts: variety and lexical density. A written and spoken corpus of an international study was used, grouped in four different age groups. For each group, both lexical measures were calculated. They found that the density measure for spoken texts is not much different between the texts of adults and young people, but in written texts they can identify a difference between age groups. However, the measure of variety allows to better differentiate the age groups than density, since adults have a rich and varied vocabulary.

Another study [58] analyzed the relationship between lexical variation and sophistication measures with oral proficiency in L2 learners. The main conclusion was that helping learners to increase their knowledge of less-commonly-used words will impact positively on their lexical variation and the overall lexical richness. Furthermore, these measures are related to the features that academic advisors review in a research proposal, for example, the repetition of words.

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

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 in the case of graduate texts provide a guideline to be used as a reference, *i.e*. establish a scale to evaluate new undergraduate level drafts. For each section, a scale with the 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 dimension in every section, for instance, for the objective section, results in density, variety and sophistication dimension of the graduate corpus (60 samples) were averaged. Consequently, different ranges defining the scale of the seven sections of a draft were obtained.

Finally, a web interface was designed so that students could be evaluated based on the three dimensions of the proposal draft and improve it if a result of low lexical richness was obtained. This interface was applied in the Pilot Testing described in section 5.2.1.

4.1.4. Coherence method

A coherence method was designed to assessment the seven sections of a proposal draft. Below, such method is described:

The global coherence refers to the thematic similarity between the section under evaluation and the semantic space, mined from the corpus in the domain of computer science and information technologies, described in 5.1. For example, if the text under evaluation contains concepts thematically close to biology, their measure of coherence will be poor, since our corpus is of the computer and information technologies domain. Figure 7 depicts the concept of global coherence [27].

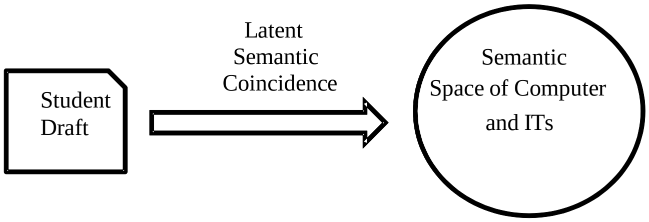

Student Draft
Latent Semantic Coincidence
Semantic Space of Computer and ITs

Figure 7. Global Coherence

Under the concept of global coherence, a method was designed which includes the following components:

Knowledge Source Corpus (described in 5.1): the first step was to gather documents in Spanish, such as student theses and research proposals, previously reviewed and approved. Both kinds of documents came from undergraduate and graduate level. With this corpus, the semantic spaces were extracted for each section, *i.e.* there were seven corpora to mine (see Figure 8).

Semantic Space (defined in 2.1.3): to extract the semantic space, terms of the input elements of a proposal draft were truncated (stemmed). Images, tables and figures were ignored. The goal of the stemming process is to reduce the variations of each word. For example, words "computer" and "computers" (in Spanish computadora, computadoras) are similar, so the process would produce a word stem "comput". The Freeling tool was used for stemming. In this way, many related terms were grouped, reducing the number of terms for building the semantic space.

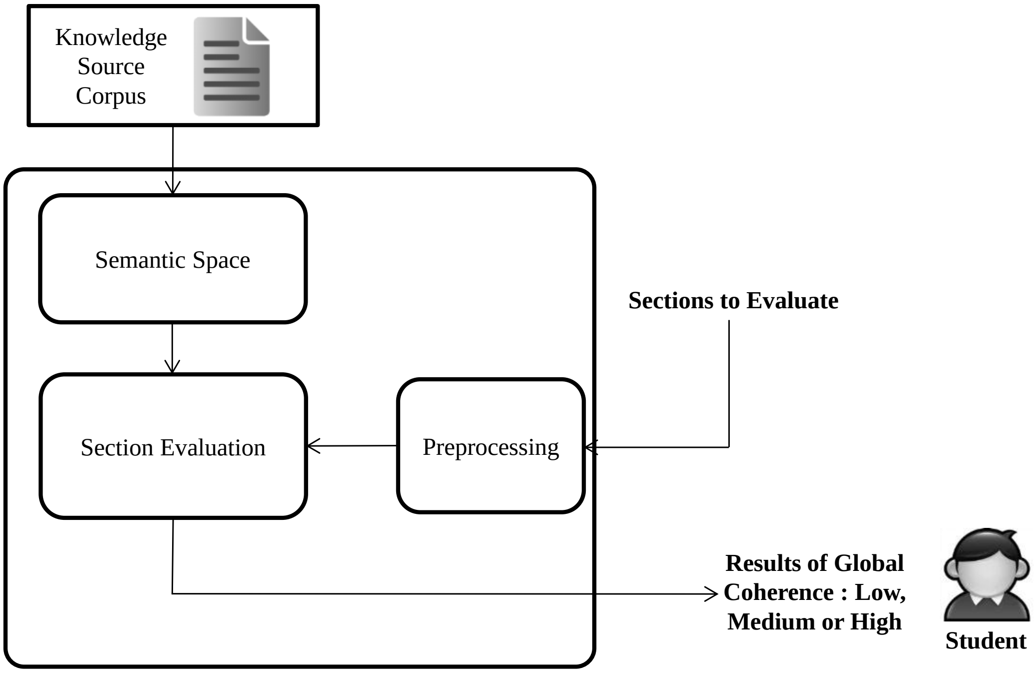


Figure 8. Coherence evaluation method

Afterward, each corpus of the sections was processed by removing stopwords such as articles, prepositions, pronouns, conjunctions, etc. for instance, "of", "the", "by" (in Spanish *de*, *la*, *por*). These stop words were supplied by NLTK-Snowball[9].

Having the vocabulary of each section, a term-document matrix was built. This matrix was processed to compute weights according to *tf-idf*, where *tf* represents the absolute frequency of appearance of a term in a document, and *idf* is the inverse frequency of the term in the documents of the collection, *i.e*. the weight of a term in a document increases if

[9] http://www.nltk.org/

it occurs frequently in such document and decreases if it appears in many (most) of the documents.

Afterwards, the LSA technique was applied to obtain the semantic space of our corpus. LSA technique was selected taking into account the correlation analysis conducted by Lapata and Barzilay [38], in which this technique presents an acceptable level. In this analysis, the correlation between humans and the different techniques was computed after assessing 90 summaries in English language. The techniques evaluated by the researchers were: Entity Grid, Overlap, LSA, HStO, Lesk, Jcon, Lin, and Resnik. The first three techniques with good performance were:

- HStO: it uses WordNet for identify the relations (antonymy, meronymy, hyponymy).
- Entity Grid: it is used to evaluate local coherence and
- LSA: it is based on word co-occurrence, also, in [35] was applied to texts in Spanish (news).

Considering the results obtained by Lapata and Barzilay, we decided to use the LSA technique because it did not require the use of WordNet tool.

Sections to Evaluate: these correspond to the sections that the student wants to evaluate, so they were analyzed one a time (*i.e.* there is no need to parse sections). The method allows evaluation of seven sections of a proposal: problem statement, justification, objective, research questions, hypothesis, methodology, and conclusions.

Preprocessing: this part of the model considers the stemming, stop word removal and computation of the tf-idf weights on the section to be evaluated. Once these processes have been applied, the text is ready to compare against the corresponding semantic space to measure similarity.

Section Evaluation: the section under evaluation is compared against the corresponding semantic space. For this purpose, the cosine similarity measure was applied to the input vectors obtained from the section and those vectors coding the semantic space.

According to this expression, the similarity is one when the angle between the two vectors is 0 degrees, that is, the vectors are pointing in the same direction and are parallel. This result expresses the highest coherence in the text. It get 0 when the vectors are orthogonal, corresponding to null coherence.

Results of Global Coherence: instead of reporting a numerical value as result of the coherence evaluation, the result is expressed in terms of three levels: High, Medium and Low. To achieve this qualitative scale of coherence, a process was applied, setting thresholds to determine each level. This information was obtained taking as reference the graduate corpus described in section 5.1, under the premise that the level of graduate students writing is better than those at undergrad level.

An experiment was set to validate the proposed model, involving human reviewers to compare the results of the method against their grades and calculating an agreement measure. Particularly, an agreement was computed in terms of Fleiss and Cohen Kappa previously defined in section 2.4.

All the collection was sent to three instructors for evaluation. They had previous experience in advising students in the preparation of drafts in the fields of computing and information technologies. 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 high level meant that the text had a strong coherence or relationship to the domain of computing and information technologies and the low level meant that the relationship was 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.*

One can observe that words “systems” and “software” are very close to the domain, including the term “architecture” in the context of the previous terms fit within the domain of computing. Likewise, words with less thematic load such as “development” or “analyze” are often used in the domain.

Low Coherence: *Identify feedback effect 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 are barely used in the domain of interest.

The assessment led to the examples exclusion rated as low by at least two reviewers and works showing no agreement since they will bias the construction of the semantic spaces.

For instance, if an objective was labeled as High by two or three of the human reviewers, this example will be part of the training corpus since, according to reviewers, such objective was indeed highly coherent with the domain. In case that only one reviewer assigned High grade, this objective will not be part of the training corpus since there is a doubt about its coherence, and it introduces noise into the corresponding semantic space.

On the other hand, the assessments on the test set made possible comparing the automatic evaluation of coherence, after extracting the semantic space and defining a grading scale. Once instructors had evaluated the whole collection (training and test subsets) the level of agreement was evaluated.

The thresholds for levels High, Medium and Low in the system were established using as a basis the average obtained when evaluating the training corpus (elements labeled with a high level) with a cross validation, *i.e.* one element of the training corpus was removed from the corpus and the semantic space was generated with the remaining examples.

Then, the standard deviation of the obtained values was calculated, and the high level was calculated as the average plus one sigma and low as the average as minus one sigma. Previously, the normality of the data was corroborated, with 95% of confidence. With the use of one sigma for thresholds, it can be ensured that the results will be in a close range to the average obtained with the best documents (labeled as high). In this case if the result is closest to the upper limit, it means that the text is closer to the domain of computing and shows strong evidence of global coherence.

Also having the semantic spaces for the different sections of our mining subset, then one can evaluate automatically the corresponding section in the test subset. Then, there are elements to evaluate the level of agreement between the grade assigned by the system and by instructors.

4.1.5. Conceptual Flow

Conceptual flow solution incorporates different schemes and each is applied depending on the section that is being assessed. A document with an appropriate structure presents a clear flow of topics through their paragraphs. For example, in the five paragraph essay paradigm proposed by [48], introduction and conclusion share the main topic, this is the theme or subject matter of the essay. The remaining paragraphs in such approach named “body paragraphs” contain details of the essay argumentation and are linked via the main topic. This approach is similar to the proposal drafts, for example in a conclusion section, the paragraphs are connected by the same main topic, and also contains paragraphs that support the results, considering the topic of the problem statement.

Figure 9 depicts the conceptual flow method. First the preprocessing section is done in two main tasks. The first focuses on segmenting each section (Justification, Problem Statement and Conclusions) into paragraphs, *i.e*. sequences of sentences bound by line feeds. Entity Grid (EGrid) tool requires as input, the text in a Treebank format. The second task is a translation from Spanish to English, since our corpus is in Spanish. The result of the translation enables to process the text with an English parser, in particular it was used Stanford (Currently, the parsers for Spanish do not adhere to the Treebank tags).

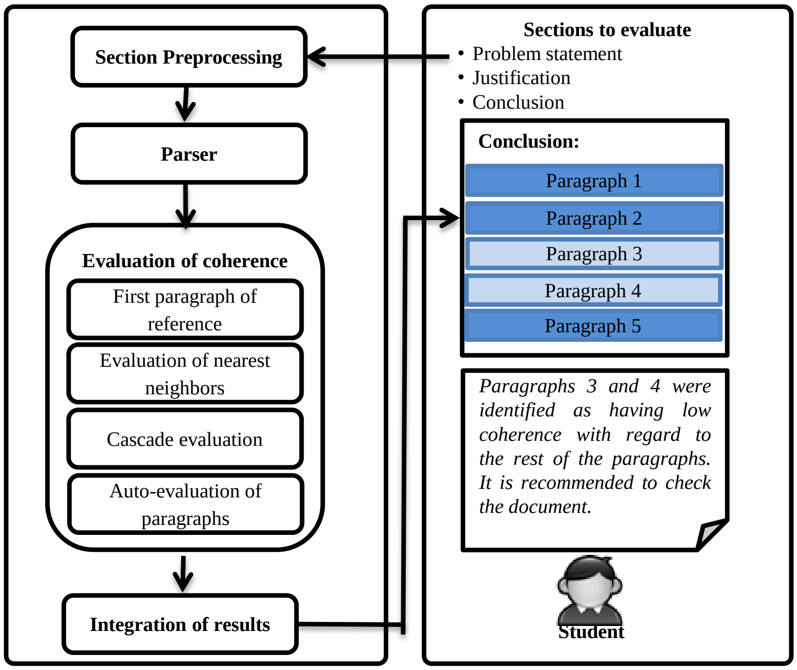
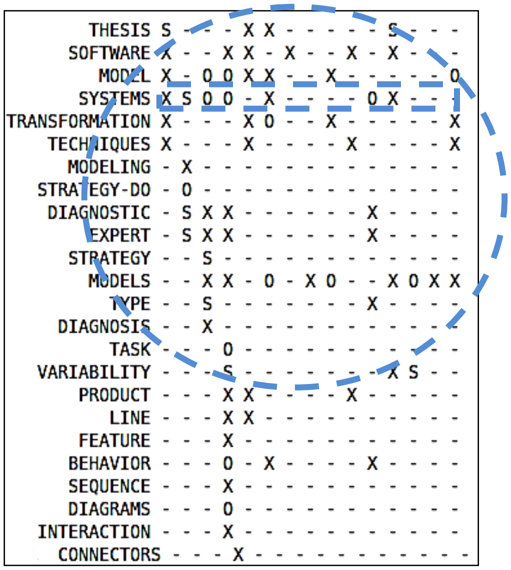

Section Preprocessing
Parser
Evaluation of coherence
First paragraph of reference
Evaluation of nearest neighbors
Cascade evaluation
Auto-evaluation of paragraphs
Integration of results
Sections to evaluate
• Problem statement
• Justification
• Conclusion
Conclusion:
Paragraph 1
Paragraph 2
Paragraph 3
Paragraph 4
Paragraph 5
Paragraphs 3 and 4 were identified as having low coherence with regard to the rest of the paragraphs. It is recommended to check the document.
Student

Figure 9. Conceptual flow method

The schemes for analysis emerged after analyzing the behavior of the transitions on the EGrid [18] of 10% of samples of the three sections of the graduate corpus. The analysis was done employing the Coherence Toolkit, using basically the commands Train and DiscriminateRand (DR) [49]. The first command generates the models, and the second uses the generated model and evaluates the paragraphs. The second command performs a binary discrimination task, which tests the ability of the generated model to differentiate between a document in its original order and a random permutation of that document, and produces results in terms of Accuracy and F-measure. So, it was generated the model of a paragraph and then applied DR to evaluate a paragraph of the same document with that model. The idea was that the model could predict whether the paragraph was related to the model: the higher the F-measure and accuracy, the stronger the relationship, *i.e*. there is evidence that the two paragraphs have a flow of conceptual sequence. Otherwise, the evaluated paragraph is not connected. Each of the schemes are now explained.

4.1.5.1. First Paragraph of Reference (FPR)

When analyzing Conclusions, it was observed that most of the transitions appear as first paragraph entities, *i.e*. entities identified by the tool in the first paragraph are further shown in the rest. Figure 10 depicts that the entities identified in the first paragraph are within the dotted circle, the rest corresponding to the second paragraph. In addition, the remaining paragraphs also included same entities as the first. Note that the transitions appeared in a sequence [S, O, X, -]. Subject (S), object (O), and neither (X), i.e. transitions between subject and object means that the concept is present throughout the text. For instance, the term "system" appears in the second sentence as *subject*, later the same term appears as an *object*. Later, in the third and fourth sentences it appeared as the object, and finally was identified as object in the eleventh sentence. These transitions provide evidence that most paragraphs are adequately connected in term of concepts. Sentences are represented vertically and horizontally the concepts (see Figure 10). In the section 2.2 the Entity Grid technique was described.



```
        THESIS S - - - X X - - - - - S - - -
      SOFTWARE X - - X X - X - - X - X - - -
         MODEL X - O O X X - - X - - - - - O
       SYSTEMS X S O O - X - - - - O X - - -
TRANSFORMATION X - - - X O - - X - - - - - X
    TECHNIQUES X - - - X - - - - X - - - - X
      MODELING - X - - - - - - - - - - - - -
   STRATEGY-DO - O - - - - - - - - - - - - -
    DIAGNOSTIC - S X X - - - - - - X - - - -
        EXPERT - S X X - - - - - - X - - - -
      STRATEGY - - S - - - - - - - - - - - -
        MODELS - - X X - O - X O - - X O X X
          TYPE - - S - - - - - - - X - - - -
     DIAGNOSIS - - X - - - - - - - - - - - -
          TASK - - - O - - - - - - - - - - -
   VARIABILITY - - - S - - - - - - - X S - -
       PRODUCT - - - X X - - - - X - - - - -
          LINE - - - X X - - - - - - - - - -
       FEATURE - - - X - - - - - - - - - - -
      BEHAVIOR - - - O - X - - - - X - - - -
      SEQUENCE - - - X - - - - - - - - - - -
      DIAGRAMS - - - O - - - - - - - - - - -
   INTERACTION - - - X - - - - - - - - - - -
     CONNECTORS - - - X - - - - - - - - - - -
```

Figure 10. EGrid of a Conclusion

The FPR scheme begins by generating a model of the first paragraph. Then, subsequent paragraphs are evaluated with FPR model, expecting that they get a value near to 1 (applying the command DiscriminateRand). The results provided by the tool are in the range from 0 to 1, where zero indicates a random flow of conceptual sequence, in some cases a null flow. A result of one implies the existence of a relationship between the model and the evaluated text, in this case a subsequent paragraph, *i.e*. the flow of conceptual sequences is strong.

After comparing each result obtained by the EGrid of the graduate corpus with the content of each Conclusions section, it was found a strong flow to those paragraphs that showed a value higher than 0.5 and a weak flow to those below that value. In assessing the subsequent paragraphs, it was expected the results to be above zero, which would show that the section had a fit conceptual sequence.

Table 13 shows how a conclusion is analyzed. First, the conclusions were assessed with the FPR scheme, generating the first model of paragraph 1, which was used to evaluate the remaining paragraphs.

Table 13. Conclusion section of undergrad corpus

<table>
<tr><th colspan="8">Paragraphs</th></tr>
<tr><td>1</td><td colspan="2">2</td><td colspan="2">3</td><td colspan="2">4</td><td rowspan="2">Method</td></tr>
<tr><td rowspan="3">Paragraph of Model 1</td><td>ACC</td><td>FM</td><td>ACC</td><td>FM</td><td>ACC</td><td>FM</td></tr>
<tr><td>0</td><td>0</td><td>0.5</td><td>0.66</td><td>0</td><td>0</td><td>FPR</td></tr>
<tr><td>0.5</td><td>0.66</td><td colspan="2">Model 2</td><td>0.9</td><td>0.9</td><td>ENN</td></tr>
</table>

ACC=Accuracy and FM=F-measure

As a result, it was found that paragraph 2 and 4 showed a null connection to the first one (referred hereon as null paragraphs). Instead, paragraph 3 displayed a good connection to the first, getting a value of 0.5 in accuracy and 0.66 in F-measure. From these preliminary results, it was assumed that the conclusion was partially disconnected.

Afterwards, it was required to relate the null paragraphs with its prior and subsequent neighbors with the ENN scheme (detailed below). In this case, there was the option to build the model of paragraph 3. As a result of evaluating paragraph 2 with paragraph 3 (model 2), a value of 0.66 was obtained in F-measure, which indicates that there is a relationship between them. Later, paragraph 4 was evaluated with model 2, finding an F-measure of 0.92 which indicates a strong connection between paragraph 3 and 4 (Table 13). Light gray shading indicates that a close left neighbor has been found. From these results, the evaluated section shows evidence of a fit conceptual sequence. In the case where a close neighbor could not be found, it would be inferred that the conclusion is not properly connected and has to be restructured.

4.1.5.2. Evaluation of Nearest Neighbors (ENN)

Is a scheme designed to evaluate null paragraphs identified after being examined with the FPR scheme. So, its main purpose is to relate null paragraphs with their prior or subsequent neighbor.

The first step is to evaluate the null paragraph with the prior neighbor paragraph model. If there is no relationship, one proceeds to evaluate the paragraph with the subsequent paragraph model. Finding a relationship between the null paragraph and its neighbor paragraphs, a connection was detected with the rest of the conclusions. Also, all paragraphs of the evaluated section would show some flow of concepts (see Figure 11).

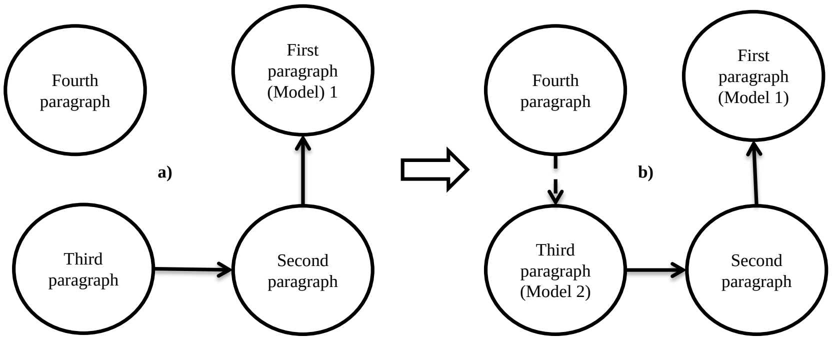


Figure 11. a) FPR results; b) ENN results

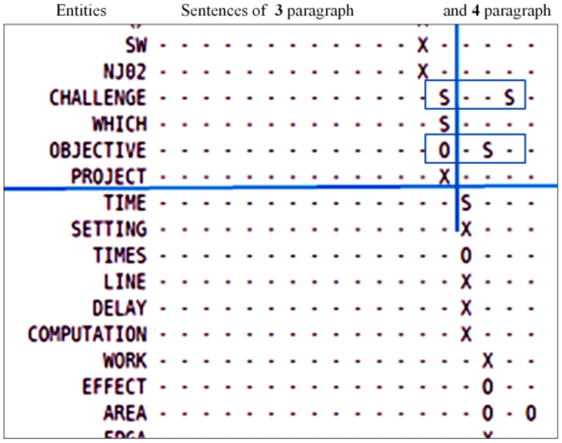
Figure 11a shows four paragraphs of a Conclusions section evaluated by FPR, with the first three paragraphs having an acceptable flow (strong connecting lines). The fourth paragraph is null and shows no connection. When applying the ENN scheme on the null paragraph, a connection to its close prior neighbor was obtained (dotted line, Figure 11b), enabling the interconnection of the entire section and the null paragraph is actually connected.

Table 13 also illustrates the application of ENN. Notice that it is not necessary to use the ENN in paragraphs where the results of the evaluation were higher than zero, since they are already showing a relationship with the first.

4.1.5.3. Cascade Evaluation (CE)

CE scheme evolved from what was observed in the EGrid regarding the Justification section. EGrid shows that transitions are distributed and do not concentrate on one position. A comparative review between the EGrid and the evaluated original text of the Justification allowed finding that some paragraphs presented a sequentially thematic relationship.

For instance, Figure 12 shows a partial EGrid of the third and fourth paragraphs of a Justification of a grad text. The third paragraph contained the entities "challenge" and "objective", which are identified as subject and object, respectively. Later, the same entities appear with subject roles in the fourth paragraph. This similarity in the entities revealed that the paragraphs are indeed directly related.



```
     Entities         Sentences of 3 paragraph          and 4 paragraph
         SW - - - - - - - - - - - - X - - - - -
       NJ02 - - - - - - - - - - - - X - - - - -
  CHALLENGE - - - - - - - - - - - - - S - - S -
      WHICH - - - - - - - - - - - - - S - - - -
  OBJECTIVE - - - - - - - - - - - - - O - S - -
    PROJECT - - - - - - - - - - - - - X - - - -
       TIME - - - - - - - - - - - - - - S - - -
    SETTING - - - - - - - - - - - - - - X - - -
      TIMES - - - - - - - - - - - - - - O - - -
       LINE - - - - - - - - - - - - - - X - - -
      DELAY - - - - - - - - - - - - - - X - - -
COMPUTATION - - - - - - - - - - - - - - X - - -
       WORK - - - - - - - - - - - - - - - X - -
     EFFECT - - - - - - - - - - - - - - - O - -
       AREA - - - - - - - - - - - - - - - O - O
```

Figure 12. EGrid of a Justification

This behavior looks like a thematic window between two paragraphs, showing a relationship to the preceding paragraph. This window moves between the rests of paragraphs. The CE scheme works by generating the model of the first paragraph and this model was used to evaluate the second paragraph. Subsequently, a model of the second paragraph was generated, and the third paragraph was evaluated with this model. This process was repeated for all paragraphs of the Justification.

An example of the undergraduate corpus evaluation of the Justification section is shown in Table 14 where the models appear as a stairway. Paragraph 2 shows a value of 0.55 for F-measure when evaluated with the model of the first paragraph, and paragraph 5 has a relationship to paragraph 4 with a value of 0.66.

Afterwards the ENN scheme was applied, but without finding any neighbor paragraph. Finally, the AP method (below) was not applied because the number of sentences was not enough. This result allowed to identify that in the middle of the Justification there was a null paragraph, *i.e*. disconnected from the remaining text.

Table 14. Justification section of undergrad corpus

<table>
<tr><th colspan="10">Paragraphs</th></tr>
<tr><td>1</td><td colspan="2">2</td><td colspan="2">3</td><td colspan="2">4</td><td colspan="2">5</td><td rowspan="2">Method</td></tr>
<tr><td>Model 1</td><td>ACC</td><td>FM</td><td>ACC</td><td>FM</td><td>ACC</td><td>FM</td><td>ACC</td><td>FM</td></tr>
<tr><td rowspan="5"></td><td>0.55</td><td>0.55</td><td colspan="2"></td><td></td><td></td><td></td><td></td><td rowspan="3">CE</td></tr>
<tr><td colspan="2">Model 2</td><td>0</td><td>0</td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2" rowspan="3"></td><td colspan="2">Model 3</td><td>0.9</td><td>0.92</td><td></td><td></td></tr>
<tr><td colspan="2">not found a close neighbor</td><td colspan="2">Model 4</td><td>0.5</td><td>0.66</td><td>ENN</td></tr>
<tr><td colspan="2">Paragraph with 2 sentences</td><td colspan="4"></td><td>AP</td></tr>
</table>

ACC=Accuracy and FM = F-measure

4.1.5.4. Auto-evaluation of paragraphs (AP)

Was designed to evaluate null paragraphs that remained after being evaluated with any of the previous schemes. It was decided to assess the paragraphs individually, hoping that they were at least connected properly within, even though they were not related to the other paragraphs. The first step was to divide each of the paragraphs in two parts, but only those paragraphs with at least four sentences, since the tool does not generate models for one sentence. Afterwards, the FPR method was applied, *i.e*. generate the model of the first "paragraph" and then the second is evaluated with this model. So, it was produced a value of connection for the individual paragraph.

4.1.6. Weak Sentences

The method has a Weak Sentences Identifier module, which contains three main components (Figure 13). Identifying Weak Sentences (IWS) is responsible for discerning whether a sentence is weak or strong. It includes five models that use different techniques and resources, such as: lexical richness, similarity between sentences measurement, use of speculative terms and overlap with terms from the conclusion sections of approved theses.

The component Classifying Weak Sentences (CWS) looks at sentences that were identified as weak by the previous component and determines the kind of weakness. This component takes advantage of a corpus tagged by human annotators to train a model with the main weaknesses identified by them. The component Customized Feedback to Students

(CFS) selects a paragraph of a conclusion section from an approved thesis to provide students support to improve their writing.

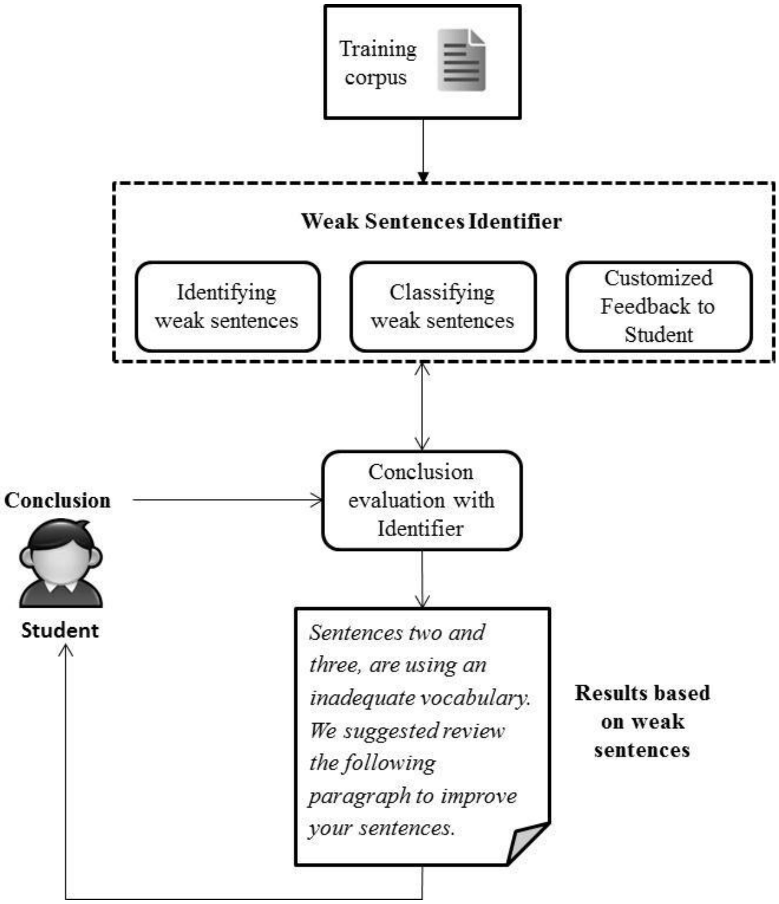

Training corpus
Weak Sentences Identifier
Identifying weak sentences
Classifying weak sentences
Customized Feedback to Student
Conclusion
Conclusion evaluation with Identifier
Student
Sentences two and three, are using an inadequate vocabulary. We suggested review the following paragraph to improve your sentences.
Results based on weak sentences

Figure 13. Weak Sentences Identifier

Identifying a weak sentence in the conclusion is complicated for human annotators, since it requires expertise in computer science thesis advising and the ability to discern if a sentence in a conclusion complies with minimum requirements. The annotators were instructors from public universities, teaching courses and advising students on their final semesters about their theses. The background of the annotators was in computer science, specifically in information technologies. Annotators tagged each sentence as strong or weak. For sentences identified as weak, annotators provided the type of weakness.

Annotators were provided with a guide defining the qualities of a good conclusion and including examples of weak and strong sentences. The criteria used by annotators to identify the two kinds of sentences, based on university institutional guides and authors of methodology books were: a global response to the research question, compliance with each of the research objectives, the acceptance or rejection of the hypothesis, and the contrast between the fundamentals and results.

4.1.6.1. Identifying Weak Sentences

The models developed seek to capture features that reflect weak or strong sentences. A sentence presents weakness when concepts are written in general instead of specific terms, or if there is absence of reflections and value judgments. These types of sentences do not fit in the conclusion section. An example of a weak sentence, part of a thesis about computer networks, states:

Security should not be a problem, neither for networks nor in everyday life, but as some humans do not have a social conscience either by greed, a bad curiosity, ambition.

One can notice in the sentence that the student is using speculative words such as "*should*". Also the student expresses a philosophic argument that fits better into an introduction section as a motivation of the problem.

A strong sentence means that the text is reasonable and makes sense in a conclusion. For instance:

The new system will help to lower costs for man/hours invested in the maintenance of the infrastructure

One can see in the sentence that the student is writing a possible consequence of the implementation of the system. This work used five models for identifying poor student sentences in the conclusion section. These are detailed next.

Lexical Richness (LR): is the first model and involves the variety, density and sophistication measures. These features were described at the first level of the analysis.

Coverage Model (CM): the second model is inspired in the Token Level Similarity approach, that it is used to resolve a basic textual entailment. The Coverage model uses the Core-Concepts (CC) obtained by MEAD[10], a tool that allows automatically extracting a set of CC from the corpus of graduate theses. CC are key ideas that support the learning of the student. In MEAD, each CC is represented by one sentence. In this work, the CC represent strong sentences drawn from high quality, graduate thesis conclusion sections that can be used to identify weak sentences in TSU conclusions. An example of core concept extracted from the corpus is:

[10] http://www.summarization.com/mead/

The Unified Modeling Language (UML) allows developers to make inroads into the paradigm of object-oriented design at analysis and design level, but not in implementation level.

One can identify that “UML” can be used with the paradigm of “object-oriented”. A TSU conclusion that references similar topics can be given the paragraph containing this CC as an example. Such feedback can help them to improve their writing.

To use CCs in finding weak sentences, a model similar to that of [50] was employed, counting the number of words in common between all CCs and the student sentence, obtaining a score. A low score means the sentence appears unlike any graduate thesis sentence and may therefore be a poor sentence. The score is calculated by dividing the intersection of the words of the student sentence and the CC domain (expanded with synonyms) by the words of the student. The result is given in a range from 0 to 1, where a value close to 0 means that sentence is far from the CCs. It was also explored a variant, Coverage Model (CM2), which eliminates the empty (stop) words when comparing TSU sentences and CCs.

Similarity of Cosine (SC): with the goal of identifying sentences that do not fit with the rest of the sentences in the conclusion section, it was computed the cosine of the sentence to be classified when compared to all sentences of the conclusion. The cosine was calculated treating each sentence as a vector of term counts. The SC score was defined as the average of all similarities calculated for the sentence to be classified. These steps were applied for each sentence. Therefore, there are averages for each sentence of the conclusion. A sentence with a low average is unlike all other sentences in its conclusion and may therefore be a weak sentence.

The last model developed addresses the use of Speculative Words (SW) in the sentences by students. It was taken as a reference the table of lexical features provided by [51] that includes modal auxiliaries (may, might, could, would, should), evidential verbs (appear, seem), adjectives (likely, probable, possible), adverbs (probably, possibly, perhaps, generally) and nouns (possibility, suggestion). The Spanish versions were used. The conclusions have to show evidence of reflections and the fulfillment of the objectives. If the sentences contain speculative words then the conclusion is anomalous. For example, the phrase “probably the results” shows that the student does not have certainty of results and this sentence may be a weak sentence. Also a variant of this model was developed: Speculative Words Expanded (SWE), which includes colloquialisms and synonyms. The idea was to identify words that do not add value to the conclusion.

4.1.6.2. Classifying the Weak Sentences

The goal of this component is to take the student sentence tagged as weak by the previous component and a variant of this model was developed: what kind of weakness the sentence shows. All the sentences tagged as weak by human annotators (165) were selected to train a model. Annotators provided a description of why they thought a sentence was weak. It was executed an unsupervised clustering over these descriptions with the objective of identifying which were the most common weaknesses of the corpus. It was applied Latent Semantic Analysis clustering. Upon manually inspecting the clusters, the following main types of weaknesses were identified:

1. The sentence was not connected to research results (**NC**): the sentence does not show evidence of some kind of analysis or reflection of the results. Also, there is an absence of arguments to provide support for the results. For instance: *The strategy that was used for this project had good results.* We can see that the sentence does not show a contrast between strategy and results.
2. The sentence is written in General Terms (**GT**): the text is like an introduction, justification or related work. The sentence does not add value to the conclusion, only extending it. For example: *This indicates that Ethernet will continue to evolve while other transmission technologies disappear*. Observe that the student fails to take a personal position.
3. The vocabulary used in the sentence is not adequate (**VO**): this includes errors such as repetition of terms or words, and terms that are distant from the topic. Example: *Now, it is time to lay hold of all the tools that exist in our environment in order to economize on operations and do not forget to seek the security of this.* The student uses informal phrases like “lay hold”.

   Models useful for classifying the different kinds of weakness are: SW, SWE, CM, CM2 and LR (described in previous section).

4.1.6.3. Customized Feedback to Students

The kind of feedback sent to students depends on the weakness type identified by the component CWS. For every weakness, the method will send a message to the student, showing as an example of a good conclusion, a paragraph from a Bachelor's or Master's degree thesis (which are of higher quality than TSU theses). The paragraph is intended to contrast with the identified problem in the TSU conclusion. For example, if the TSU conclusion sentence was classified as GT, then it means the paragraphs should not be written in general terms.

This component takes as a reference the CCs that resulted from running MEAD on the corpus of Bachelor and Master degree. We seek to identify which of the CCs are related to each of the types of weaknesses identified. Thus we have two groups of CCs: the first is related to the GT class and the second group is related to the kind of NC_VO class. Each group is ordered depending on how appropriate they are in correcting weak sentences with that class. Given a sentence that has been identified as weak and classified as NC, VO or GT, this component selects the most highly ranked CC of the appropriate class, retrieves from the corpus the paragraph containing the CC, and sends the paragraph as feedback. It is expected that the paragraph helps the students to improve their conclusions. If the paragraph is not helpful, the system can continue down the ranked list to send additional feedback.

Models employed for feedback that are most useful for finding CCs to address the different kinds of weakness are: for NC_VO class we used SW, SWE, CM, CM2 and for GT class we used SC, SW, SWE, CM, CM2, and LR (all described above).

4.1.7. Speculation, Opinion and Coverage

The analysis of conclusions section was done with a designed method that involves three features: speculation, opinion and the linking between the objective and conclusion section (Coverage).

The method has a Mining Component, which contains three main models. The Coverage model is responsible for identifying whether or not a conclusion sentence has a connection with the general objective, in terms of the main concepts. This is a way to take into account the recommendations of authors from research methodologies books. Opinion model processes each sentence to identify terms with an opinion load, evidencing the presence of opinions or value judgments formulated by the students. The idea is to help the student to undertake a process of analyzing results and that the conclusion is not just a list of achieved activities. The final Speculation model identifies if the student expressed future work, or possible derivations of his/her work. Below it is described the features evaluated by the method:

- ✓ *Coverage*: the model seeks to assess if any of the sentences of the conclusion section have some connection with the general objective.
- ✓ *Opinion*: value judgments and reflections elaborated by students are key features of a conclusion. With the proposed model in this work, the intention is to assess if the conclusion has an acceptable level of opinion.
- ✓ *Speculation*: The proposed model identifies speculative terms in conclusion sentences. As a result of the reflections of the research done, it is expected that the conclusion shows evidence of future work or possible derivations.

4.1.7.1. Coverage Model

This model seeks to identify whether or not the conclusion shows connection with the general objective. It is expected that some sentences display this relation. This task can be solved as a basic case of textual entailment. The general objective corresponds to the hypothesis and the conclusion corresponds with the text.

In the first step, empty words were removed from documents of graduate and undergraduate level, in conclusion section and general objective. Also each term was stemmed with the Freeling tool.

For the conclusion section, a group of sentences was used, while in objectives the full text was considered, that is an objective was considered as one sentence. For computing coverage, the following expression was applied:

$$Coverage(C) = \frac{\max(\#(So \cap S_i))}{N}$$

where S is a list of words of an objective (So) or a sentence *i* of conclusion (S_i), and N is the number of terms in the objective. The value of the sentence with highest coverage is kept. The result is in a range from 0 to 1, where a value close to 0 means that sentence is far from the objective.

4.1.7.2. Opinion Model

The goal of this model is to identify if the conclusion section shows evidence of opinions. For example:

It was demonstrated that the use of conceptual graphs and general semantic representations in text mining is feasible, especially beneficial for improving the descriptive level results.

One can observe that terms as feasible and beneficial imply an opinion. To take into account terms that reflect an opinion or value judgments, it was employed SentiWordNet, a lexical resource for English, which associates an opinion score to each term depending of the sense (e.g. noun, adjective), with three numerical values for objectivity, subjectivity and neutrality (each between 0 and 1). Each conclusion was translated to English employing Google Translator [56], and then, empty words were removed and the value for each sentence was computed, searching each term in SentiWordNet 3.0 (considering all senses of the words). For instance, the Opinion load measures (non null) in the conclusion given above:

S2: Possible(0.37) make(0.13) communication(0.04) minimize energy(0.21) use(0.07) common(0.29). Total = 1.11

The term possible presents a 0.37 opinion load, this result is computed regarding the average of all opinion loads (as a noun has 2 senses and an adjective has 2 senses). The total displayed is the sum of all terms. It was foreseen that in conclusion (S2+S3) an acceptable load was expressed.

4.1.7.3. Speculation Model

The model identifies evidence of sentences that describe future work or derivations of the research. For this purpose, two lists of speculative terms were merged. The first list includes lexical features provided by [51] that include modal auxiliaries, epistemic verbs, adjectives, adverbs and nouns. The second list, the “Bioscope corpus”, consists of three parts, namely medical free texts (radiology reports), biological full papers and biological scientific abstracts. The dataset contains annotations at the token level for negative and speculative keywords [52], tagged by two independent linguists following guidelines. To obtain this list, the terms tagged as the speculation type were extracted from the XML file (e.g. the terms *suggesting* and *could*):

<cue type=”speculation” ref=”X1.6.2”>suggesting</cue>

<cue type=”speculation” ref=”X1.7.1”>could</cue>

After the extraction of speculation terms, the two lists were combined, with the goal of gathering a more complete list. Terms that appear in both lists were weighted by 2 and those terms that only appear in a list were given the value of 1. Weighted terms indicate higher speculation in the sentences. Each term of the merged list was translated to Spanish, producing a list of 227 speculative terms.

4.1.8. Methodological Questions

A method was developed to identify answers to methodological questions within the general objective. The following steps describe the method:

The first step was the construction of a language model using text segments marked by the annotators corresponding to answers to the methodological question *What will you do?* (text-1Q). For the experiment it was taken into account:

1. The segments of text-1Q where 3 or 4 annotators had agreed, ignoring the rest. The objectives that met the criteria were 232 out of a total of 300.
2. The training group was established as 80% (186) of text-1Q segments.
3. The test group was the remaining 20% (46)
4. The elements of each group were selected randomly, and then agreements between annotators of the training and test group were computed. It was obtained 0.597 of agreement for training group that corresponds to the “Moderate” level. For the test group, it was gotten a “Substantial” level (0.624).

An example of the text-1Q used for training is:

Propose VP a DT methodology NN based VBN on IN interaction NN patterns NNS

The sample contains the token + grammatical class. It is noteworthy, that some segments of text-1Q had a longer extension than the example. Afterward, the objectives of the test group were evaluated. The aim was to identify the text-1Q in the objective of test sample.

The methodology described above defines the steps in the experiments developed in Chapter 5. In each experiment, details are given of corpus used and the results achieved.

Chapter V: Experiments and Analysis

5. Experiments and Analysis

Experiments and results are presented according to the four levels outlined in the proposed solution: Lexicon, Coherence, Language Models and Methodological questions. Each of the experiments included the analysis and discussion of the results.

For the experiments, a corpus described in section 5.1 was collected allowing the characterization of the thesis elements. In addition to that, a percentage of the corpus elements was used to perform validation tasks, according to each experiment. The collected corpus is composed of theses and research proposals of graduate and undergraduate levels. It is worth mentioning that the corpus was increasing during the research.

This chapter describes the experiments performed on each of the levels of the developed solution described in Chapter 4. First the experiments related with lexical richness are detailed. Afterward the results achieved in the level of coherence are described. Later the experiments of the third level are presented: Weak Sentences and Automatic Assessment of Students Texts. Finally, the results achieved on level fourth which corresponds to the methodological questions are described.

5.1. Text Corpus and Human Reviewers

The collection was integrated by 468 documents: theses and students' research proposals written in Spanish language. Each item of the collected corpus is a document that was evaluated at some point by a reviewing committee. In the corpus, two kinds of students can be distinguished, graduate level (Doctoral and Master degree) and Undergraduate (Bachelor and TSU degree). Documents come from universities and research centers in Mexico.

To build the collection, complete documents (theses and research proposals) were downloaded from universities' public repositories[11]. Subsequently, from each document the following elements were extracted by hand: Problem Statement, Justification, Research Questions, Hypothesis, Objectives, Methodology and Conclusions section. The following table shows the amount of collected data per study degree.

[11] http://bc.unam.mx/index-alterno.html
[11] http://www.dirbibliotecas.ipn.mx/Paginas/Tesis_Electronicas.aspx
[11] http://catarina.udlap.mx/u_dl_a/tales/
[11] http://www.remeri.org.mx/repositorios/

Table 15. The Corpus

Degree	Items
Doctoral	59
Master	181
Bachelor	150
TSU	78
Total	468

It can be seen that Doctoral and TSU degrees have fewer documents compared to Master and Bachelor degrees. Regarding Doctoral degree, the theses production level is lower in comparison to other study levels. There are fewer available repositories for TSU degree documents. However, when the elements of graduate (240) and undergraduate level (228) were counted, both groups were balanced. This corpus was built during the development of this thesis, therefore, in the first experiments fewer elements were used in relation to the content presented in Table 15.

Below, Table 16 details each element extracted from the corpus. The amount displayed on Table 12 totals 2,216 elements extracted from the whole corpus. Note that the Title element was not processed in this work, however, it was extracted for future work purposes.

Table 16. Detailed Corpus

Corpus			
Section	Graduate	Undergraduate	Total
Problem statement	132	100	232
Justification	108	132	240
Research Questions	133	19	152
Hypothesis	71	26	97
Objectives	218	212	430
Methodology	113	71	184
Conclusions	216	197	413
Title	240	228	468
Total	1231	985	2216

The graduate level has 1231 elements and the undergraduate 985. It is relevant mentioning that some elements were not found corpus documents; only the title was extracted from every document.

5.1.1. Human reviewers

For validation, a set of elements from the corpus was selected for each experiment. In the following subsections, each experiment defines its own set of validation. The developed methods are supported by the Fleiss and Cohen Kappa Test. A group of human reviewers were employed for the tagging process.

Human reviewers were in charge of tagging the validation sets. Thus, a reference set (gold standard) could be obtained. The reviewers are professors from public universities in computing field. They have expertise in reviewing theses. Finally, the collected corpus was made available to the community on the website: www.coltypi.org. The full corpus was published with an option to download elements of a given thesis.

5.2. First level: Lexical Analysis

These experiments were focused on examining lexical richness in documents written in Spanish, at graduate and undergraduate levels. In this section the results achieved by methodology to assessment lexical richness are described. Also, it is included a section with the results of correlation analysis among the measures assessed. In addition, it is detailed a pilot testing conducted with undergraduate students.

5.2.1. Lexical Methodology

The results obtained in this section confirm the expectation that the graduate level works had higher lexical richness on each dimension in the objective, hypothesis and research questions sections. However, concerning the rest of the sections of the thesis, the richness results of both graduate and undergraduate thesis are close. But, a correlation analysis between dimensions makes evident that graduate students have better writing skills in relation to their lexical competence. The experiments were developed using the methodology described previously in Chapter 4.

5.2.1.1. Corpus

A corpus of different elements was gathered –as it was previously explained- within proposal documents written in Spanish. The corpus consists of a total of 410 collected samples, as detailed in Table 17. The first kind of texts includes documents of Doctoral and Master level (PG). The second kind comprises documents of Bachelor (BA) and Advanced

College-level Technician degree (TSU). The corpus domain is computing and information technologies.

Table 17. Spanish Text Corpus for Lexical Analysis

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

5.2.1.2. Experimental Results

Results were divided into two groups, considering the size of the sections. The first block includes sections with short texts (*i.e.* objectives, questions and hypothesis) and the second with long texts (problem, justification, methodology and conclusions). Research questions and Hypotheses are not present on TSU degree works because those sections are not required at this level. In the first block, documents of graduate level scored better in every dimension (see Table 15, Lexical Richness).

Regarding objective section, TSU degree texts 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 dimensions of each evaluated level, evidence was found about a correlation arising from undergraduate and graduate level between LV and LS; for research questions section (PG and BA level), the data were significant at 0.01 level (p-value <0.01), this means that the appearance of more content words probably was reflected as sophisticated words in the text. Lexical density and variety do not show dependence between them because correlation was low and p-value was not significant (above 0.01). The same results were obtained for density and sophistication dimensions. (Table 18, Correlations).

Table 18. Lexical richness and correlations: first block

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

It was observed in the objectives section of TSU documents that the correlation is strong between LV-LS (the data was significant at the 0.01 level), 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 could cause a low level of lexical richness compared to the high level, at least three of the four results with significance.

In the second block, the results show a slight variation from the first block. One can notice here 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 19). This dimension can be used to differentiate the levels. Moreover, the graduate level can be used as reference, since it showed the highest sophistication.

In the conclusions section, lexical richness in undergraduate texts 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 as much as the average employed by the undergraduate level (BA and TSU).

When performing correlation analysis on the second block sections, negative values were observed in the justification and conclusion sections. However, only in the conclusions section (LV-LS), the results were significant at 0.01 level.

Table 19. Lexical richness and correlations: second block.

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

Negative correlation indicates that as a dimension increases the other decreases. In this case, variety is higher than the sophistication. Also, considering texts sizes and having lower density, it is likely to find more content words, which could be sophisticated words 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, not displaying sufficient arguments (after a manual review of these conclusions), as expected or suggested by research methodology authors [3]. Therefore, the undergraduate level could not be considered better than the graduate level. Finally, Figure 8 depicts the average of the three measures obtained for both subsets of the corpus.

As expected, one can notice that graduate documents produced higher averages than undergrad drafts, for the different elements (sections). It can be observed that the first three blocks achieved greater lexical richness in comparison to the second block. Short texts are less likely to contain repeated terms.

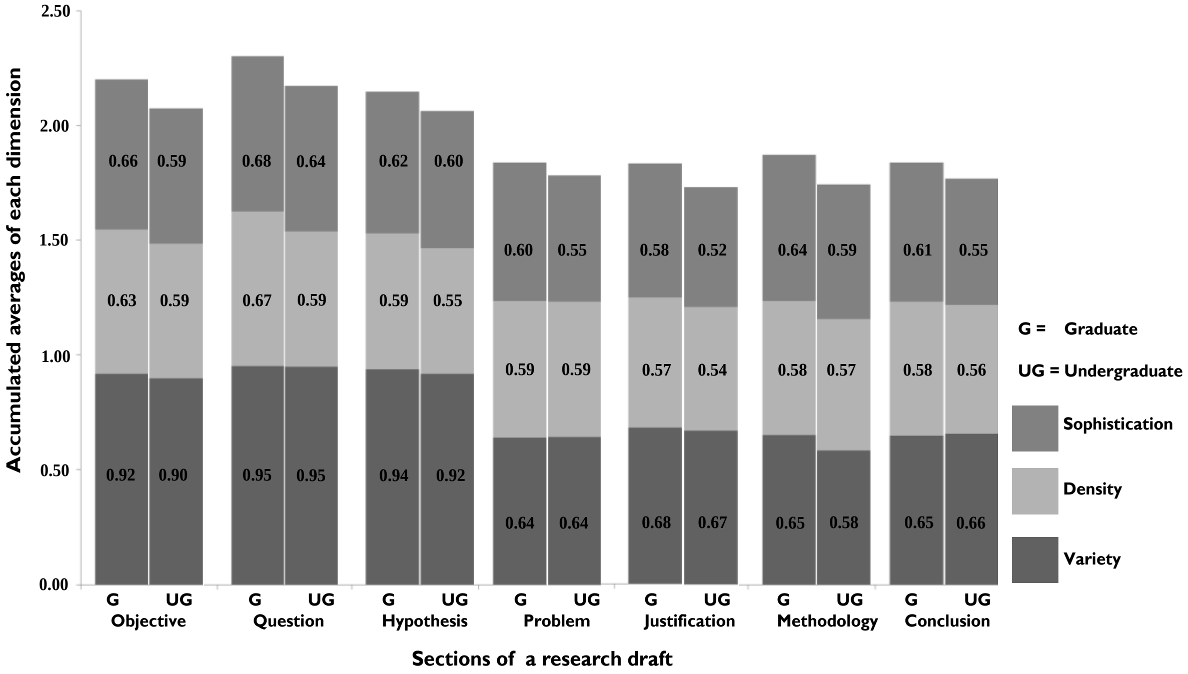


Figure 14. Lexical analysis results

The section that reaches the highest value of lexical richness is the research questions with 2.3 and the section with lowest lexical richness is justification with 1.83, both sections of graduate level. In the undergraduate level works, the same sections scored 2.17 and 1.73, respectively. These results provide evidence that students are more concrete on short sections at both levels. It can be seen that difference between the highest and the lowest of lexical richness for graduate level is 0.47 and for undergraduate level is 0.44. Also, standard deviation for undergraduate level is below that of graduate level. This leads us to assume that despite having less lexical richness at undergrad level, sections are more homogeneous than those of graduate level. When ordering sections at both levels, from highest to lowest for lexical richness, practically similar behavior was found between both. (see Table 20).

It can be asserted that students are following a similar writing process, so their deficiencies in the three dimensions evaluated could be related. Educational Institutions can formulate similar strategies to improve both lexical richness levels in sections that are located on the same level.

Table 20. Lexical richness orderly of highest to lower

Undergraduate	Lexical Richness	Graduate	Lexical Richness
Questions	2.17	Questions	2.30
Objectives	2.07	Objectives	2.20
Hypothesis	2.06	Hypothesis	2.15
Problem	1.78	Methodology	1.87
Conclusions	1.77	Conclusions	1.84
Methodology	1.74	Problem	1.84
Justification	1.73	Justification	1.83

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

(1) Objective (High level): *Implement an algorithm based on hierarchical structures with volume enveloping of spheres for collision detection*[12].

(2) 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*[13].

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 variety. A qualitative analysis was also performed to compare results produced by the lexical methodology and manual review on each section by an academic advisor. This analysis was to verify that there exists congruence between them. The feedback provided by the lexical methodology is defined by academic advisors and, depending on evaluation results, suggestions and tips are given to help the student improve the lexical richness.

5.2.1.3. Pilot Testing

A pilot testing was performed to assess the impact/benefit of using an intelligent tutor focused on lexical richness in elements of a research project. This experiment involved

[12] Translation of original title in Spanish: "Implementar un algoritmo basado en estructuras jerárquicas con volúmenes envolventes de esferas para la detección de colisiones."

[13] Translation of original title in Spanish: "Crear un sistema de administración de información de franquicias con datos relevantes de cada establecimiento y datos del personal de cada franquicia así como también referencias de los franquiciatarios y personal de confianza que manejan las franquicias."

undergraduate students of Telematics Engineering and Systems from Universidad de la Sierra, in Mexico.

Two groups, each integrated by 14 students were considered: the experimental and the control. All students were enrolled on their seventh semester and were taking the Simulation course in which they were required to prepare a research project proposal.

Both groups received instructions on how to write a problem statement, justification, hypothesis and general objective. Students were informed about the proposal requirements concerning variety and usage of jargon from the domain of computer science and information technologies. Moreover, they should not use many stopwords (prepositions, articles, and so on) and should also avoid the abuse of terms of content, such as using certain words repeatedly. Neither group of the pilot testing received information of how to compute each measure.

The control group had a traditional monitor, *i.e.* an academic advisor reviewing their documents, while the experimental group had access to the lexical methodology (built-in an intelligent tutor[14]) 24 hours a day. All documents produced by both groups were evaluated with the lexical methodology to compare results of lexical richness among them. The foremost hypothesis to be validated in this testing pilot was: “The use of the lexical methodology allows students from the experimental group to generate documents with better parameters in terms of richness in comparison to documents produced by the control group.”

Figure 15 depicts the average of the three measures obtained for both groups on the pilot testing. One can notice that the experimental group produced higher averages than control group, for the different elements (sections). The section that reached the highest value of lexical richness was Hypothesis with 2.24 and the section with lowest lexical richness was Problem Statement with 2.04, both from the experimental group.

In the control group, values were 2.01 and 1.85 in Objective and Problem Statement, respectively. These results provide evidence that students are more concrete on short sections at both levels, as in Objective and Hypothesis sections. Also, the standard deviation of three measures for control group was 0.17, while the experimental group was

[14] The lexical methodology was embedded in an Intelligent Tutor (IT). This IT, was done by a collaborator of these experiments, see other publications section.

0.14. This could indicate that students in the experimental group have a more homogeneous writing style, regarding the three evaluated measures.

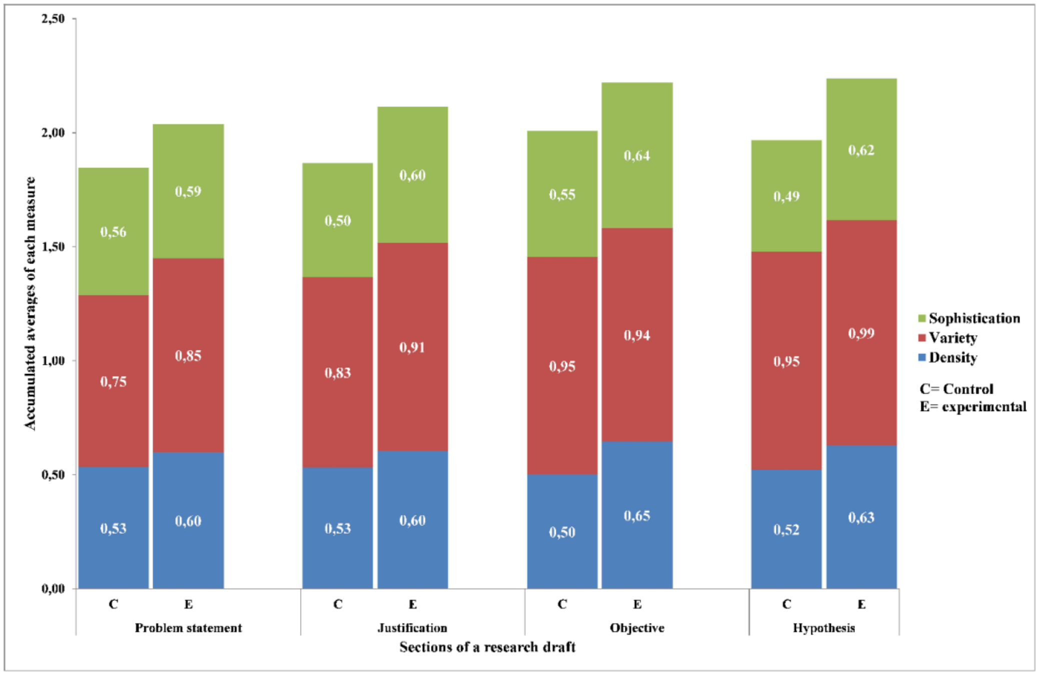


Figure 15. Lexical Analysis Results of Pilot Test

Also, a hypothesis test was applied to two independent samples with different standard deviations in order to validate results. The null hypothesis and the alternative hypothesis are shown below. The confidence level was 95%.

$$H_0 : \mu_{Experimental} \leq \mu_{Contrast}$$

$$H_1 : \mu_{Experimental} > \mu_{Contrast}$$

Hypothesis tests were performed to each section and to each measure. Table 21 shows results regarding hypothesis H_0. One can notice that in density, the null hypothesis was rejected in the four evaluated sections. This result shows statistic evidence that the intelligent tutor and the lexical methodology allows support students in these sections.

Table 21. Results of Statistical Analysis of Pilot Test

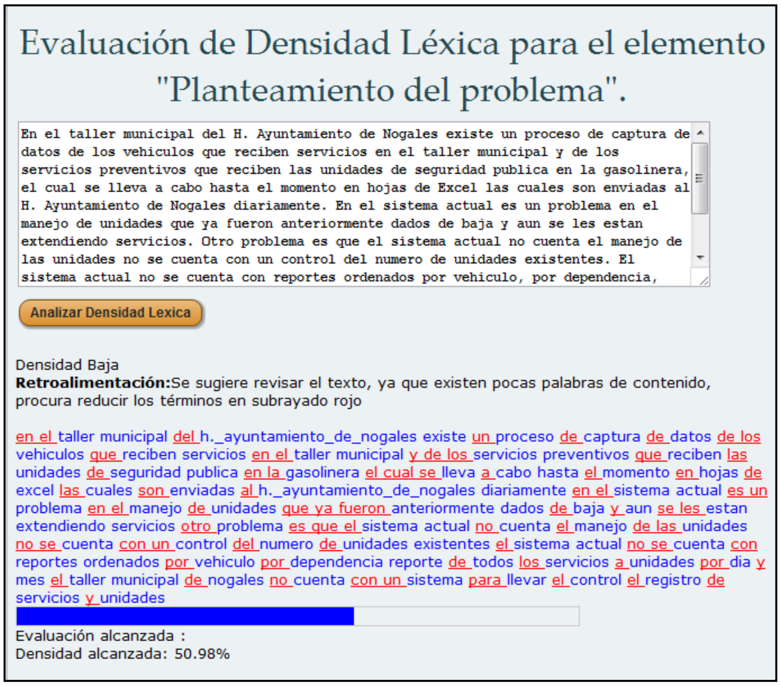
Evaluated section	Measures		
	Density	Variety	Sophistication
Problem Statement	Rejected	Rejected	No Rejected
Justification	Rejected	Rejected	Rejected
Objective	Rejected	No Rejected	No Rejected
Hypothesis	Rejected	Rejected	Rejected

The null hypothesis was not rejected in lexical variety of the Objective, this situation probably is because students have more experience with this section. It is common for a student to write life objective or a specific project objective, while other sections are less common. Finally, concerning the sophistication measure in the Problem Statement and Objective sections, the null hypothesis was not rejected; whereas the Justification and Hypothesis sections had opposite results. It is interesting to mention that the Objective section, the null hypothesis is rejected in variety and sophistication measures that could indicate that students have better writing skills for this section.

From this statistical analysis, it can be stated that the intelligent tutor and the lexical methodology helped undergraduate students improve on three lexical aspects: variety, density and sophistication. It is noteworthy that sophistication is a plus in student writing.

Below, Figure 16. depicts the output of the evaluation of the lexical density. The feedback provided by the analyzer is defined by academic advisors and, depending on the results of the evaluation, suggestions and tips are given to help the student improve the lexical richness.

Figure 16. Output screen for measuring density (in Spanish).



Evaluación de Densidad Léxica para el elemento "Planteamiento del problema".

```
En el taller municipal del H. Ayuntamiento de Nogales existe un proceso de captura de
datos de los vehiculos que reciben servicios en el taller municipal y de los
servicios preventivos que reciben las unidades de seguridad publica en la gasolinera,
el cual se lleva a cabo hasta el momento en hojas de Excel las cuales son enviadas al
H. Ayuntamiento de Nogales diariamente. En el sistema actual es un problema en el
manejo de unidades que ya fueron anteriormente dados de baja y aun se les estan
extendiendo servicios. Otro problema es que el sistema actual no cuenta el manejo de
las unidades no se cuenta con un control del numero de unidades existentes. El
sistema actual no se cuenta con reportes ordenados por vehiculo, por dependencia,
```

Densidad Baja
Retroalimentación: Se sugiere revisar el texto, ya que existen pocas palabras de contenido, procura reducir los términos en subrayado rojo

en el taller municipal del h._ayuntamiento_de_nogales existe un proceso de captura de datos de los vehiculos que reciben servicios en el taller municipal y de los servicios preventivos que reciben las unidades de seguridad publica en la gasolinera el cual se lleva a cabo hasta el momento en hojas de excel las cuales son enviadas al h._ayuntamiento_de_nogales diariamente en el sistema actual es un problema en el manejo de unidades que ya fueron anteriormente dados de baja y aun se les estan extendiendo servicios otro problema es que el sistema actual no cuenta el manejo de las unidades no se cuenta con un control del numero de unidades existentes el sistema actual no se cuenta con reportes ordenados por vehiculo por dependencia reporte de todos los servicios a unidades por dia y mes el taller municipal de nogales no cuenta con un sistema para llevar el control el registro de servicios y unidades

Evaluación alcanzada :
Densidad alcanzada: 50.98%

The feedback of the lexical density analyzer and the classification assigned to the statement problem proposed by the student is Low Density due to the large number of stopwords relative to content words, and the IT sends a message to the student with a feedback according to the level assigned. The message displayed is “we suggested reviewing the text, there are few content words, seeks to reduce the terms outlined in red” in the paragraph analyzed. We observe that empty words are underlined to indicate the student have to try to replace (reduce) them, and the interface includes a progress bar to indicate graphically the progress of this writing, in this case a 50.98% of advance.

5.2.1.4. Discussion

It can be asserted that graduate students from the information technologies area have better writing skills. These results fostered the development of a web tool where students can analyze the vocabulary of each of the essential sections of their proposal drafts.

An interesting finding was that dimension which mostly differentiates between them 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 (*i.e.* longer sections). This result does not imply a successful result because it would be necessary to verify that the student really does argue properly on each of the sections of a proposal draft, as suggested by authors of research methodology.

Using lexical methodology for research project drafts aims to support teachers in reviewing research proposals providing material to the student, by tracking their progress and lexically analyzing the drafting of their writings. The pilot test done with two groups of students provided some evidence that the use of lexical methodology helped students improve their writings in terms of their lexical richness.

These results must be qualitatively validated with instructors once additional features get included. This lexical methodology is the first step of evaluation of the student paper. After the document is evaluated by the lexical methodology, the document will be assessed at the following levels of model solution of this thesis. It is possible that in this first level the student reaches a high lexical level, however this result does not mean that the document does not require further review. The solution to four stages allows the document will be polished in the following stages.

The lexical methodology reported here for Spanish is not difficult to move to a different language, given that it only depends on two language resources: the stop and common word lists.

5.3. Second level: Coherence

This section describes the Global Coherence method and the results from the validation experiments. The Latent Semantic Analysis technique described in Chapter 2 was employed on the corpora of research proposals and theses to further assess proposal drafts of college students in information technologies and computer science.

The connection between paragraphs involves the interconnection of each of the sentences within the paragraph through its grammar constituents as subject and object. These constituents are observed as a pattern and allow to correctly interpreting the information in the text [37]. Also, results are presented regarding local coherence from a conceptual flow approach into paragraphs.

5.3.1. Global Coherence

The experiments were done on graduate and undergraduate corpora to validate the process and the experiments involved human reviewers to compare the results of the method with those of the reviewers, so the agreement measures were computed. The method was described in Chapter 4.

5.3.1.1. Corpus

The whole corpus consists of a total of 410 collected samples, and was tagged by annotators indicating the level of coherence (High, Medium and Low). Elements with High level of coherence were used to build the semantic space.

Table 22. Training and Test Corpus

Sections	Training	Test	Tagged as High Level
Problem statement	40	14	23
Justification	40	18	20
Objectives	60	20	40
Research questions	40	10	36
Hypothesis	40	20	20
Methodology	40	14	27
Conclusions	40	14	24

5.3.1.2. Experimental results

In this section, the results obtained for each section evaluated by the method of coherence method are presented. In addition, the results of agreement among annotators and the method are included.

5.3.1.3. Objective

The Fleiss Kappa coefficient of agreement was computed for the three reviewers considering the test corpus. Table 23 shows the Fleiss Kappa results for each level, for the objective section. The reviewers had a Substantial agreement for the Low and High grades, and a Poor agreement in Medium grades.

Table 23. Kappa for Test Corpus (Objective)

Levels	Reviewers	Method vs. Reviewers
High	0.6862	0.0000
Medium	-0.0378	0.2609
Low	0.7353	0.4218
Overall	0.5458	0.2237

For the results obtained, it could be concluded that reviewers clearly identified High and Low levels but not those in the middle. The overall level achieved between evaluators was 0.54, this corresponds to a Moderate confidence of agreement for the experiment

These levels let automating the evaluation of the coherence method. In particular, for the objective section, it was obtained 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 was defined, the test samples were evaluated 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 method results. Table 23, also shows the Cohen's Kappa results for the human versus coherence method, being Fair and Moderate for Medium and Low levels, with a Fair overall agreement. In addition, despite that the High level does not reach an acceptable level yet, low and medium levels of coherence are already detected, giving certain confidence to the instructor that the method can identify objectives with deficiencies.

5.3.1.4. Problem Statement

For this section, the level of agreement of the three reviewers was very low (High=0.22, Medium=0.18 and Low=0) and only two of them assigned high level grades (35% first reviewer and 83% second reviewer). Therefore, it was decided to consider only two reviewers in the experiment. The high level grades would be used for mining.

The overall level achieved between evaluators was 0.68, this giving Substantial confidence of agreement for the experiment. These levels allow automating the evaluation of the coherence method. For this section, after getting the semantic space, it was gotten a low average of 0.127 with and standard deviation of 0.057, leading to setting the thresholds at 0.07 for Low and 0.18 for High.

Table 24. Kappa for Test Corpus (Problem)

Levels	Reviewers	Method vs. Reviewers
High	1.000	1.0000
Medium	1.000	0.3300
Low	0.000	0.0000
Overall	0.680	0.4000

As observed in the Kappa-Cohen values between method and reviewers, there is a Perfect level of agreement in High grades but a margin for improvement in the Medium grade since this is Fair as the overall agreement. Since human reviewers did not agree on tagging problem statements with a low grade in the test set, no agreements with the method could be expected. But, to find out if the approach can identify the low grades, examples labeled as Low were taken from the graduate corpus. These examples were not included in the training set, but for exploration purpose, the examples were evaluated with the coherence method and add to previous results obtained with test set. With these results, the Cohen Kappa between human reviewers and the method was computed. According to the results, the Kappa showed an improvement for low and medium level. High level maintained the level of agreement, the medium and low level of Fair changed to Moderate, with 0.43 and 0.40 respectively. The overall agreement level was 0.49 which represents a Moderate level.

5.3.1.5. Hypothesis

As it was previously explained, only two of the three human reviewers were considered. Kappa results between human reviewers were Acceptable with 0.301, similarly as the

Kappa between the method and human reviewers was Acceptable with 0.2558 (see Table 25). However, it was lower than in the objective and problem statement sections. With the purpose of defining the evaluation scale an average of 0.636 was gotten with a standard deviation of 0.236, resulting in the high threshold of 0.87 and the low threshold at 0.4.

Table 25. Kappa for test corpus (Hypothesis)

Levels	Reviewers	Method vs. Reviewers
High	0.3953	0.5294
Medium	0.2528	0.1428
Low	0.0000	0.0000
Overall	0.3010	0.2558

For the High level the Kappa value among human reviewers and the method was 0.5294 corresponding to *Moderate* according to Kappa scale. The zero value of agreement among human reviewers affects the outcome of the method to the low level. Although only examples with High level were used to define the scale of three levels, also, the human reviewer's distribution was unbalanced.

Low grades in Hypothesis presented a similar complication as the Problem Statement section, where human reviewers did not agree tagging examples with low grade. Again, to find out if the approach could identify low grades, the examples labeled as Low in the graduate corpus were considered.

Moreover, the examples were evaluated with the coherence method and added to the previous results that were obtained with the test set. When executing Cohen Kappa between human reviewers and the method, the values high, medium and low were 0.6363, 0.111 and 0.333, respectively. It was observed that Kappa for High level is "Substantial". The medium level remains at "Slight" level and the Low moved from "Poor" to "Fair". The overall level of agreement was "Fair". In this case, despite that the medium grade did not reach an acceptable level, the low level reached an acceptable agreement. The method can give certain confidence to the instructor that a hypothesis with deficiencies will be identified by the system, and then the advisor can suggest students to improve their Hypotheses.

5.3.1.6. Justification

The kappa values achieved were lower compared to the previous sections, even so the level is Acceptable or Fair. For the justification section, after computing the semantic space, it was obtained an average of 0.137 with a standard deviation of 0.066 leading to set the thresholds at 0.07 for Low and 0.2 for High.

An Acceptable level was obtained between the reviewers and the method with 0.39 (Table 26). Moreover, high level had a Moderate agreement and the medium level was Acceptable. Observe that the levels of agreement between human reviewers were Fair, despite having a balanced assignment of grades. The reason could be that the high grade was assigned with a similar percentage but not to the same samples.

Table 26. Kappa for test corpus (Justification)

Levels	Reviewers	Method vs. Reviewers
High	0.2200	0.5800
Medium	0.2075	0.3600
Low	0.2758	0.0000
Overall	0.2283	0.3900

Unlike the previous two sections, in this section the human reviewers tagged some examples with low grade in the test set, showing a Fair agreement in terms of kappa value. A strategy implemented to raise the agreement results for low grades was using half sigma to define the thresholds. The results improved for low level, but affected the medium level. The kappa values for the High and Low level were 0.33 and zero respectively.

Another attempted alternative to improve results (medium and low levels) was training a classifier (Naive Bayes), using as input vector the LSA value provided by the semantic space and the grade (class) assigned by the reviewers. As training examples, it was used the set of graduate and undergraduate texts, evaluated as low and medium. After training, the classifier had a precision of 0.714 and recall of 0.5 for the low grade. The medium level reached a precision of 0.706 and recall of 0.857. The level of agreement was Acceptable in terms of kappa. These results indicate that the classifier is a promising alternative to predict medium and low grades for this section.

5.3.1.7. Conclusions

In this section, it was obtained an average of 0.268 with a sigma of 0.247 allowing to set the thresholds for Low at 0.021 and for High level at 0.514. The level of agreement between reviewers was 0.310, corresponding to the Acceptable level. Also there was a 0.166 level of agreement among human reviewers and the method, this means a Slight level of agreement. High and medium grades were Acceptable according to a kappa of 0.280. The value of agreement was zero for low grade. This was probably due to the low coincidence of examples labeled as high. As observed in results of previous sections, the coherence method results regarding human agreement levels are close, indicating that the method is directly dependent on the level of agreement between humans.

In addition, the kappa level between human reviewers for low level was null, since none of the examples was graded as low (see Table 27). But to know if the approach could identify low grades, some examples labeled as Low were taken from the graduate corpus. The result again was unfavorable, since the values were low, according to previous values. Subsequently, in order to improve results, the classifier (Naïve Bayes) was used. For training, examples from the graduated corpus tagged as medium and low were employed. After training, the values of precision and recall were produced.

Table 27. Kappa for test corpus (conclusions)

Levels	Reviewers	Method vs. Reviewers
High	0.2857	0.2857
Medium	0.4000	0.2857
Low	0.0000	0.0000
Overall	0.3103	0.1666

The results were favorable, reaching a precision value of 1 and recall of 0.556 for the medium class, while for the low class reaching a precision of 0.556 and recall of 1. Kappa value was of 0.447, higher than using thresholds. These results indicate that for this section, the use of a classifier for predicting medium and low class seems more promising than using the average and standard deviation to define the scale. The classifier was trained with medium and low classes, since the method was built with the high class.

5.3.1.8. Research Questions

As observed in Table 28, the Medium grade level had a zero percent agreement, which it was expected since the level of agreement was very uneven between reviewers. For high grade, reviewers reached a value of 0.50 and for low grade, reviewers obtained a kappa of 0.46, which corresponds to a Moderate agreement. The threshold was set considering the average of 0.432 with a sigma of 0.286, allowing to set the Low Level at 0 .227 and the High level at 0.638, for this section. The agreement results between human reviewers and the coherence method were 0.33 for High and Low grades. This corresponds to an Acceptable level according to the range of kappa. One can notice clearly that the reviewers and our method identified High and Low grades.

Table 28. Kappa for test corpus (Questions)

Levels	Reviewers	Method vs. Reviewers
High	0.5000	0.3333
Medium	-0.0230	0.0000
Low	0.4666	0.3333
Overall	0.2727	0.2000

5.3.1.9. Methodology

The Fleiss Kappa for High level was 0.1923 between reviewers, *i.e.* Slight agreement. An average= 0.315 with a standard deviation of 0.158 allowed to set the Low grade level at 0.156 and the High grade level at 0.474, for automating the grades of the method for this section. Among human reviewers and the method, a value of 0.12 of agreement was obtained for High grades. Both values are in poor performance based on Kappa. One possible cause is that the undergrad methodologies tend to have fewer steps and a simpler elaboration than graduate level methodologies.

For Low grade, the agreement amounts to zero (Table 29). It was not possible to approach this section as a classification task since one of the reviewers did not tag low grades and the rest of the reviewers did not coincide on their grades. One possible cause of this can be the variety in writing in this section, that favored a disagreement between human reviewers.

Table 29. Kappa for test corpus (Methodology)

Levels	Reviewers (Fleiss)	Method vs. Reviewers
High	0.1923	0.1250
Medium	0.1900	0.2750
Low	-0.0500	0.0000
Overall	0.1250	0.1764

5.3.1.10.Discussion

It was 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 method was Fair. It can be concluded that the method of analysis of coherence predicts in an acceptable manner the level of global coherence, taking into account the results obtained by the method and the annotators.

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

It can be observed that if the system does not achieve at this time a higher level of agreement in the High grade level, this is not a problem since the method is being strict to assign the high level. In the experiment, the method evaluated as Medium the few highest levels assigned by the reviewers. If the method behaves more flexible and allows high level to sections 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, it was noted that between the coherence method and human evaluators, the agreement is Moderate for low levels, bringing confidence that the method is identifying those sections that were classified as Low by reviewers. After assessing coherence, the method as part of a system, can trigger feedback to the student for any of the seven selected sections in the draft.

5.3.2. Conceptual Flow

At the same level of global coherence model, the method of conceptual flow was placed. In this experiment it was explored the relationship between paragraphs in Justification, Problem Statement and Conclusion sections, this was called conceptual flow, which conveys an implicit local coherence. This section presents the results of conceptual flow analysis, using the method described in Chapter 4. This method is based on a local coherence technique called Entity Grid, previously described [37].

5.3.2.1. Corpus

The corpora consist of 240 collected samples, 120 samples for graduate (G) and 120 for undergraduate (U) level, with 40 samples of each of the sections: problem statement, justification, and conclusion. The second kind included documents of Bachelor and Advanced College-level Technician degree.

5.3.2.2. Experimental results

The objective of the experiments was to apply the method and the previously designed evaluation schemes on the corpora. In this way, it was generated a diagnosis of both levels in the Problem Statements, Justifications and Conclusions. In the experiments, the method was applied with the different schemes included. The remaining 90% of the corpus was used in each section.

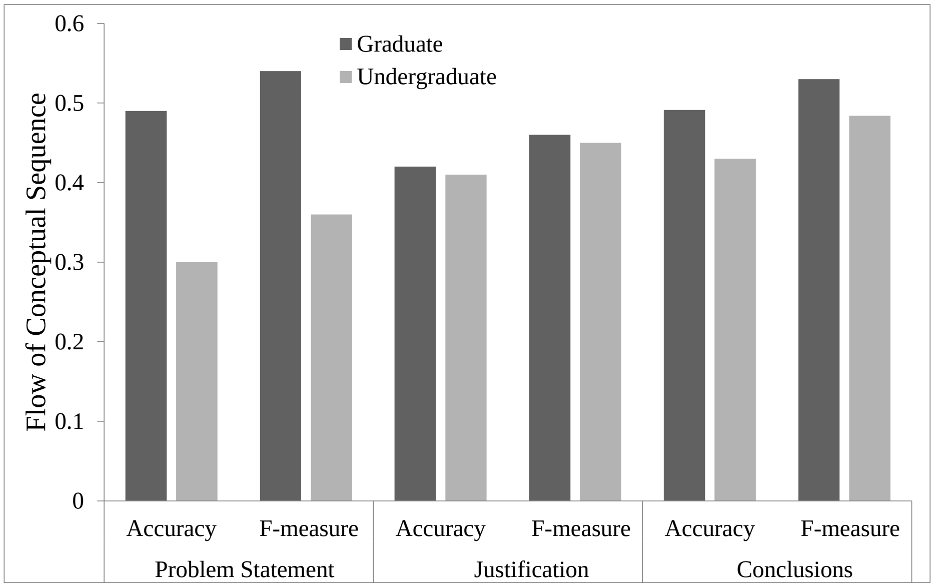


Figure 17. Results of Experimentation

Figure 17 summarizes the results in terms of average Accuracy and F-Measure values provided by the tool, for the three sections. These results show that grad students' paragraphs from the corpus are better linked than those of the undergrad students, being wider the difference in Problem Statements and closer in Justification. The results were qualitatively validated in the original text, *i.e.* it was looked for the relation of paragraphs identified by the schemes in the conclusions.

5.3.2.3. Evaluation of Schemes

Furthermore, a qualitative analysis was conducted on the examples of corpus and it was observed that the schemes allow identifying topic changes within the section evaluated. For example, if a conclusion had a low value of connectivity, this meant that there was evidence that the conclusion contained several topics, instead if a conclusion showed high connectivity the conclusion was more homogeneous regarding the main topic.

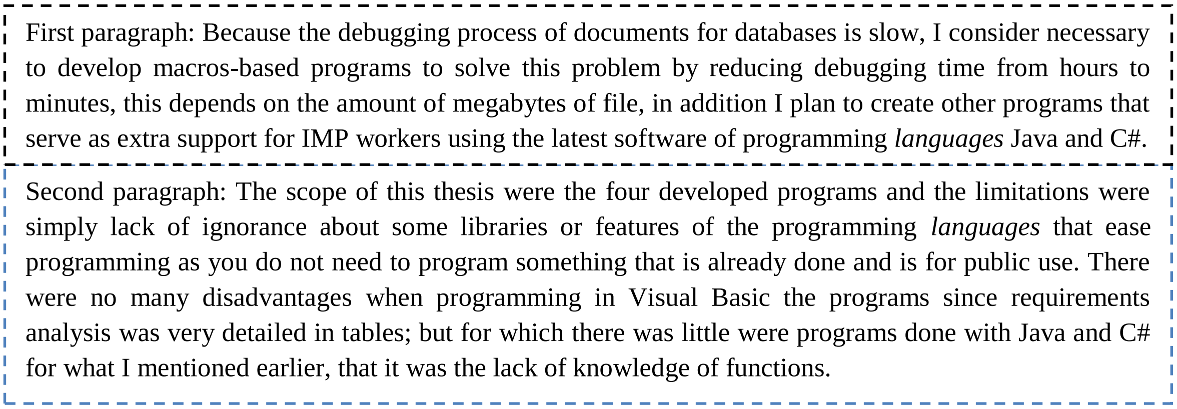
Table 30. Results of evaluation

Section	Correspondence	Non Correspondence
Problem Statement	0.64	0.36
Justification	0.70	0.30
Conclusion	0.83	0.17

Thus, it was decided to compare our schemes against a Bayesian segmentation approach, which was driven with a focus of lexical cohesion [53]. The segmentation task divides a text into a linear sequence of topically coherent segments. The authors argue that well-formed texts induce to consistent lexical distributions. The whole corpus was considered for evaluation (*i.e.* 240 samples). Each element of the corpus was evaluated using both techniques. If the methods identified the same topics that segmentation algorithm did, then the result was tagged as having a Correspondence; otherwise the result was Non Correspondence. Table 30 summarizes the percentages of examples for each case, agreeing mostly in the three sections, with higher agreement in *Conclusions*.

Below, it is shown an example of *Justification* section after the application of both techniques (the segmentation and FPR scheme). One can observe in Figure 18 that the segmentation technique identifies two topics. The first topic written by the student is a short justification related with the debugging of databases and describes the features of the debugging process (first paragraph), while the second topic presents the scopes and limitations of the task. Also the student shows details of the disadvantages of programming (second paragraph). The two topics identified used similar terms, but the association of terms is different. For instance the term “language” is used in the two segments identified, but the terms associated with them are different.

Paragraphs were taken from the source document (thesis), and the scheme FPR was applied. The first step was to generate the model of the first paragraph and then evaluate the second paragraph in the model. A null relationship was found *i.e.* the identified entities (subject and object) in the first paragraph, were not found in the second paragraph. This represents a disconnection between paragraphs, indicating that each paragraph represents different topics, coinciding with the result produced by the segmentation approach.



First paragraph: Because the debugging process of documents for databases is slow, I consider necessary to develop macros-based programs to solve this problem by reducing debugging time from hours to minutes, this depends on the amount of megabytes of file, in addition I plan to create other programs that serve as extra support for IMP workers using the latest software of programming *languages* Java and C#.

Second paragraph: The scope of this thesis were the four developed programs and the limitations were simply lack of ignorance about some libraries or features of the programming *languages* that ease programming as you do not need to program something that is already done and is for public use. There were no many disadvantages when programming in Visual Basic the programs since requirements analysis was very detailed in tables; but for which there was little were programs done with Java and C# for what I mentioned earlier, that it was the lack of knowledge of functions.

Figure 18. Justification section-undergraduate level

In [40] O'Rourke and Calvo evaluated the flow of paragraphs in university essays (English language) using two techniques MFN and SVD, described in chapter 2. The essays were classified into two groups, group A correspond to essays with score between 60 and 70 and group B with scores among 70 and 100. The authors found that the test group B obtained greater semantic flow than group A. However the measure effect size, which determines the strength of the difference between two groups, was 0.18. When calculating the effect size for the methods developed in this thesis, a value of 0.59 among graduate and undergraduate levels was obtained. This result means that the methods are able to differentiate the two levels. Besides, it was a good result considering the result of the work [40].

5.3.2.4. Discussion

Assessing the flow of concepts in proposal drafts is a complex task for computers, and sometimes even for humans. It was understood that the behavior of transitions in the Conclusions adhere to a pattern where most of the central entities concentrated in the beginning of the Grid, *i.e.* the first paragraph contains information that was further developed in the other paragraphs. This behavior corresponds to a pattern[15], which begins with the restatement of the main premise, then a summarization of the key points, and finally the formulation of recommendations, assessments and forecasts, as expressed in some academic writing guidelines. The similarity between this pattern and that observed in the EGrid was verified by reviewing the text of the Conclusions of our corpus: it was observed that when the student redefined the main problem, he/she used many of the key terms of his proposal, which were reflected in the subsequent paragraphs.

[15] http://learninghub.une.edu.au/tlc/aso/aso-online/academic-writing/

Problem statement section showed a similar pattern as Conclusions. However, a slight difference in some cases was that most of entities appeared in the first two paragraphs. Regarding Justification section, it was seen that there was a pattern where different entities were referred in a chained way, corresponding to different issues discussed at a time, as expected for this section.

Finally, the method with their different schemes can be easily applied directly to English drafts, only by omitting the translation step that was needed for Spanish. Moreover, student drafts in other domains can also be analyzed without too much trouble.

5.4. Third level: Language Models

This section describes two analyses specific for the conclusion section. The first experiment aims to identify sentences that do not fit into a conclusion, using knowledge (core-concepts) of the corpus collected and applying learning techniques. The second experiment seeks to identify internal elements of a conclusion with automatic assessment, including the connection of the general objective with the conclusion section. Results and agreement evaluation between annotators are provided.

5.4.1. Weak Sentences in Conclusion section

For these experiments, the focus was primarily on the conclusions section of a thesis. Three components are presented: Identifying Weak Sentences, Classifying the Weak Sentences, Customizing Feedback to Students. The proposed method identifies weaknesses in sentences, such as the use of general instead of specific terms, or the absence of reflections and personal opinions. For instance:

In the project, we have developed two concept tests, one has been to do the survey that collected data from people, and another has been to make a concept test of the peripheral[16]

Here, it can be observed that the student describes a part of the experimentation, instead of providing a value judgment of the results. This sentence would be more suitable to another section of the thesis, for example, the methodology section. The method starts by identifying weak sentences in the student conclusion. If the sentence is weak, the system identifies the type of weakness. Finally, the method sends feedback to the student depending on the type of weakness that was found. Also, initial models were provided

[16] All example sentences have been translated to English from the original in Spanish

including their evaluations for each component regarding the agreement level between annotator and the method. The experimental analysis was done with the method described in the Chapter 4.

5.4.1.1. Corpus

Fifty five Advanced College-level Technician degree (TSU) level theses were gathered. Then, from the conclusion section, 544 sentences were obtained. Finally, the sentences were sent to human annotators to be tagged with strong and weak classes.

In addition to that, 210 Bachelor and Master Degree theses were collected. These documents were used by the component IWS after doing Unsupervised Clustering. The aim of the clustering was to identify sentences that are representative of concepts found in approved theses. This clustering was used only in the component IWS [55]. The component CFS also uses the corpus of Bachelor and Master degree to send suggestions for students, depending on the kind of identified weakness.

5.4.1.2. Experimental Results

The first step was submitting the sentences of the conclusions to human annotators, with the aim of generating a gold standard. The sentences were tagged with the class “weak” and “strong”, and for each “weak” sentence, a description of why it was considered “weak” was provided. A total of 165 sentences were tagged with weak class and 329 sentences with strong class. Table 31 shows the confusion matrix of agreement and disagreement between annotators. Note that most agreements were on strong sentences.

Table 31. Confusion matrix between annotators

Class	Weak	Strong
Weak	48	120
Strong	36	340

The Kappa-Cohen agreement between human annotators was of 0.25, corresponding to a Fair level of agreement. A third annotator adjudicated the disagreements between the two primary annotators.

5.4.1.3. Identifying Weak Sentences

Each sentence was processed with the models to obtain features. The results were used as input to a classifier (NaiveBayesMultinomial). Lexical Richness alone was used as baseline. Below, the results obtained by classifiers with 10-fold cross-validation are presented. Since the goal is to remedy "weak" sentences, the main interest is in the precision and recall of the "weak" class, but it is also in showing the performance of the "strong" class.

Table 32 shows that all models outperformed the baseline F-measure of 0.527. It was added to the baseline different model features with the goal of improving precision and recall. The system with the highest F-Measure for Weak sentences (0.622) was SC+SWE+CM2+LR, though SC+CM+LR achieved slightly higher precision (0.613 vs. 0.603). This may suggest that identifying speculative words generally improves recall, though at a small cost to precision.

For the task, the purpose was to find a balance between precision and recall, like that of the SC+SWE+CM2+LR model, since the system has to clearly distinguish a weak sentence in the conclusion -otherwise the system could confuse the student- and at the same time should have good coverage.

Table 32. Classifying results

Models	Precision	Recall	F-Measure	Class
LR	0.556	0.5	0.527	Weak
	0.582	0.636	0.608	Strong
SC+SW+CM+LR	0.567	0.506	0.535	Weak
	0.589	0.647	0.617	Strong
SC+CM+ LR	0.613	0.548	0.579	Weak
	0.624	0.685	0.653	Strong
SC+SW2+CM2+LR	0.603	0.643	0.622	Weak
	0.653	0.614	0.633	Strong

In the work of Bethard et al. [50] they developed an intelligent tutor to identify science concepts and student misconceptions (Elementary level). One of the methods developed by the researchers was named Identifying student misconceptions, which focused on identifying misconceptions, obtaining a MAP of 64%. The IWS method developed in this thesis reached an F- measure of 0.622.

5.4.1.4. Classifying the Weak Sentences

The models were evaluated to identify the kind of weakness using as gold standard the "weak" class. Those weaknesses were tagged with the types already identified from annotators' descriptions: NC, GT and VO. There were 113 examples of GT, 34 of NC and 18 of VO. Because of the small number of sentences with types NC and VO, they were merged into a single NC_VO class. The results (Table 33) were obtained by classifiers using 10-fold cross-validation.

Table 33. Classification results

Models	Precision	Recall	F- Measure	Class
SW+SWE+CM+CM2+Variety	0.6	0.35	0.439	NC_VO
	0.748	0.894	0.815	GT

In this experiment the best combination of features was SW+SWE+ CM+CM2+Variety. Other combinations were tested that included density and sophistication (from the LR features) and cosine similarity models, but these did not perform as well. Sentence similarity features likely contribute less to this task because they are not focused on identifying vocabulary and term-based issues. In general, Table 30 shows that while the model's predictions of GT sentences are fairly reliable, identifying NC and VO sentences is more challenging, probably due to the small amount of training data available for these classes.

5.4.1.5. Customized Feedback to Students

The evaluation of the models consisted of identifying strong sentences that contrast with weak sentences of a particular type. Classifiers were trained on the "strong" sentences plus just the "weak" sentences of a particular type. For example, the classifier was trained to respond to GT problems on the 329 "strong" sentences, plus the 113 GT sentences. Similarly, for responding to NC_VO problems, there was training on 329 "strong" sentences plus 52 NC_VO sentences. Then, these two models were applied to the sentences (CCs) from the Bachelors' and Masters' degree theses, and ranked those sentences based on the score output by the classifier.

5.4.1.6. Discussion

This task was complex for human annotators, since it requires expertise in computer science thesis advising and discerning if a sentence complies with the minimum requirements. This work profits from the different academic advisors knowledge who have annotated the corpus. A variety of different models is employed to characterize potential problem sentences in a conclusion, and use them to generate features for supervised classifiers.

The result of F-measure achieved by the method IWS, was close to the method developed in [50]. However, the level of corpus used in this thesis, could represent more complexity to IWS method identifying a weak sentence.

Also, it was found that weak and strong sentences share features, *i.e.* a weak sentence contained similar terms as a strong sentence. For instance, a weak sentence can be written in descriptive manner and adjust to another section as introduction or justification. These differences allowed the classifiers to identify patterns.

Moreover, it was identified that the weak and strong sentences, labeled by the annotators, are not affected by the writing styles of each student. For instance, the use of the active voice or passive voice in a conclusion does not affect the tagging process of sentences, since the criteria defined to label weak or strong sentences contemplated own aspects of a conclusion, such as contrasting results.

The amount of training data for low frequency weakness types needs to be increase, *i.e.* inadequate vocabulary. This would allow the method to have better coverage of the different kinds of weaknesses, and to strengthen the features of weak and strong sentences.

As a future work, a group of 212 ordered CCs was prepared and sent to human annotators to evaluate the relevance of the CC with the weaknesses, in order to validate the schemes. Furthermore, the number of annotators could be increased, to improve the level of Kappa-Cohen agreement, taking care of including annotators with a similar academic background, e.g. instructors with computer engineering degrees.

The set of components presented in this analysis could be applied to other domains, such as the identification of weak sentences in essays of students learning English. Since many of the features are language independent, it would only be necessary to make some small number of changes in the text preprocessing and to use a corpus tagged by instructors with experience in reviewing English essays.

5.4.2. Automatic Assessment of Student Texts

In the conclusion section, a discussion of the results is expected, and that the students ponder about the whole research work. In particular, a good conclusion has to include the following features: an analysis of compliance with the research objectives, a global response to the problem statement, a contrast between results and theoretical framework, future research work and acceptance or rejection of the established hypothesis [54].

5.4.2.1. Corpus

The corpus contains conclusions of graduate (Master and Doctoral degrees) and undergraduate level (Bachelor and TSU). Also, the associated general objectives were gathered from each of the conclusions. In total, there are 312 conclusions and objectives (Table 34).

Table 34. Corpus

Level	Objective-conclusions
Doctoral	26
Master	126
Bachelor	101
TSU	59

From the described corpus, 30 conclusions were selected for validation with their corresponding objectives, 15 of bachelor and 15 of TSU level. Each conclusion was tagged by two annotators. The tagging process included marking the text that reveals the presence of Coverage (gray text) and Speculation (underline text). To assess the Opinion, a scale of three levels was established ("Yes, a lot", "Yes, a little", and "No opinion"). Each of the annotators had experience in the review process of theses. For instance, sentences of an undergrad objective-conclusion pair tagged by the annotators are:

Objective:

S1: *Develop a system of monitoring control and power of light in common areas through a programmable logic controller (PLC).*

Annotated Conclusions:

S2: *It was possible to establish the communication between the software (LabVIEW) and hardware (PLC), to minimize energy used in labs, cubicles and common areas presented.*

S3: *So the power control system based on PLC presented meets the objectives as well as minimizing energy use, is user friendly and may be expanded to multiple cubicles , labs and common areas.*

Opinion level: *Yes, a little*

The Kappa agreement between annotators for Coverage element was 0.92 that corresponds to Almost perfect. For Speculation element was 0.65 that corresponds to

Substantial. For the Opinion scale, the agreement was: 0.47 (Moderate), 0.21 (Fair), and 0.44 (Moderate).

5.4.2.2. Experimental results

The results achieved in each feature evaluated using the method described in Chapter 4, are described below.

5.4.2.2.1. Coverage model

The corpus tagged by annotators was employed. It was processed Coverage of each of the objective-conclusion pair and the result was placed in a scale. To build the scale, the graduate level was used as a reference of Coverage, that is after processing each objective-conclusion pair, the average of all results was computed. However, to smooth out the scale, a group of 50 elements of bachelor level was included (selected at random). Below we show the scale:

- *Coverage >= 0.12 (Average - 1σ).* This indicates that the connection between the objective and the evaluated sentence is acceptable, otherwise is taken as an absence.
- *Coverage >= 0.41(Average + 1σ).* This corresponds to a strong connection. It is expected that sentences exceed the minimum acceptable (0.12), giving evidence that the student is properly linking the objective with the conclusion paragraphs.

Finally, after evaluation of the tagged corpus (30 objective-conclusions), the Fleiss Kappa was computed between the method and the annotators, obtaining a result of 0.799, corresponding to Substantial agreement.

5.4.2.2.2. Opinion Model

Similar to the Coverage Model, the graduate level texts were taken as a reference to define a scale. However, since there are three levels of opinion, there was no smoothing. For this element, the conclusion has to reach the average level of review (*i.e*. "Yes, a little"), this will give evidence that the student is expressing judgments and opinions. Below we show the scale:

- *Opinion <= 7.84 (Average - 1σ*), these are conclusions corresponding to the level "No Opinion".
- *7.84 < Opinion < 26.98*, these are conclusions presenting the level "Yes, a little".
- *Opinion > =26.98*, these are conclusions that correspond to the level "Yes, a lot".

Regarding the previous example, the sum of S2 and S3 (1.11+1.34=2.45) fits with *No opinion* level. This result is close to the "value" assigned by annotators (*i.e. Yes, a little*). After evaluation, the Fleiss Kappa was computed between the results of the method and annotators (30 objective-conclusions pairs). It was obtained a Fair agreement for Yes, a lot (0.30), and for *Yes, a little* (0.21). For *No opinion* level (0.46), a *Moderate* agreement was obtained.

5.4.2.2.3. Speculation Model

To compute the speculation measure, it was only counted the number of speculative terms in each sentence of the conclusion (*i.e.* a scale was not stated); only the coincidence between the text marked by the annotator and the sentence with maximum value of speculation terms were considered.

For instance (conclusion of data section):

S2: The method did not find speculative terms, neither the annotators.

S3: The annotator marked the future work; also the method identified "may" as a speculative term.

Finally, the Fleiss Kappa measure was computed between the results of the method and the annotators (30 objective-conclusions), obtaining a result of 0.887 which corresponds to *Almost Perfect* agreement.

5.4.2.3. Corpus Mined

An analysis of the whole corpus was conducted using the models described above. The goal was to identify the levels of Coverage, Opinion and Speculation in the graduate and undergraduate levels. The Coverage value is the average of the maximum values of each conclusion of the corpus. The Opinion value is the average of the sum of each conclusion. In Speculation for graduate level, the sentence with the highest speculation (average) was around three terms while the undergraduate level had around two terms.

Table 35. Corpus mined

Level	Coverage	Opinion	Speculation
Graduate	0.3	20.5	3
Undergraduate	0.2	14.5	2

One can notice that the graduate level has better values than undergraduate level (see Table 35). Besides, a significance test was performed for each measure between graduate and undergraduate level (Two-Sample T-Test. $\alpha = 0.05$). For the three features, the p-value was 0.001. These results show that graduate students connect better the conclusion with the objective and express more detail about their judgments, opinions and possible derivations.

5.4.2.4. Discussion

In these experiments, we have presented a system that uses natural language processing techniques to mine specific features of writing for the conclusion section emphasized by authors of methodology or institutional guides. We took advantage of the knowledge in the theses in the corpus, reviewed by different academic advisors, when extracting the features with different proposed models.

It was found in the three features evaluated that graduate level students texts outperformed those of undergraduate level. This behavior provides evidence that students with more practice in writing (graduate level), possess better skills.

For the Opinion feature, it was considered as future work to identify if the orientation is positive or negative and determine whether or not the stated objectives were achieved. For speculation, as future work, the purpose is to include speculative phrases. Also, it is planned to increase the number of examples of the corpus to improve the level of agreement between the system and that of the annotators, specifically for opinion feature. Moreover, there are plans to conduct a pilot test with students of TSU level, with the aim to verify if the proposed system indeed helps students to improve their writing.

5.5. Fourth level: Methodological questions

At this level the purpose is to evaluate the objectives of a research proposal draft. The aim is to identify the answers to methodological questions such as: What will you do? or How are you going to do it?. This section describes the corpus tagged by four annotators. Furthermore, the agreement evaluation between annotators is provided. This level is in progress, the analysis performed to identify the answer to the questions “What will you do?” is presented below.

5.5.1. Automatic Identification of the Answer

Defining the general objective of a thesis allows students to set the way forward for the development of the thesis. Three methodological questions raised in this thesis serve as a guide for its conception. These elements are suggested by the authors of books on research methodology [3]. Below, a general objective with marked questions is presented:

Generate a support tool for the study of algebra in the bachelor level, through a computer system that manages learning objects with IEEE-LOM associated with various topics of algebra.

- 1Q: What will you do?: *Generate a support tool*.
- 2Q: What is the purpose of doing it ?: *the study of algebra in the bachelor level*
- 3Q: How are you going to do it?: *through a computer system that manages learning objects with IEEE-LOM.*

The above example can be expressed in general terms as follows:

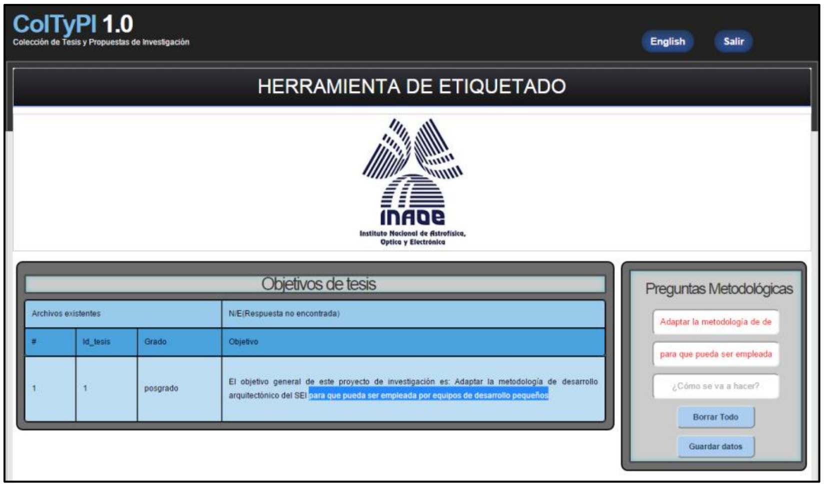
- 1Q: The *object/product* to be achieved.
- 2Q: The main purpose of the *object/product*.
- 3Q: Means (Activities, instruments) to achieve the *object/product*.

The object corresponds to the terms “*support tool*”. It is observed that the *object/product* is related to the second question since the *object* has a purpose "*the study of algebra*”. Also, the *object* is connected to the third question since activities or instruments are linked with the *object* “*through a computer system*”. The aim of this analysis is to identify what text segment contains the answer to the question "What will you do?", using language models.

5.5.1.1. Corpus

To perform the experiment, a task of tagging 300 objectives was conducted, which belong to the corpus described previously in this chapter. For this task, four annotators were selected with experience in reviewing research proposals (academic advisors of public universities). An online tool[17] was provided to annotators to perform the task. Previously, a guide with instructions for annotating was given to academic advisors. The tagging tool is part of the products obtained in this thesis (see Figure 19).

[17] www.coltypi.org


ColTyPI 1.0
Colección de Tesis y Propuestas de Investigación
English
Salir
Instituto Nacional de Astrofísica, Optica y Electrónica
Borrar Todo
Guardar datos

HERRAMIENTA DE ETIQUETADO

Objetivos de tesis

Archivos existentes			N/E(Respuesta no encontrada)
#	Id_tesis	Grado	Objetivo
1	1	posgrado	El objetivo general de este proyecto de investigación es: Adaptar la metodología de desarrollo arquitectónico del SEI para que pueda ser empleada por equipos de desarrollo pequeños

Preguntas Metodológicas

Adaptar la metodologia de de

para que pueda ser empleada

¿Cómo se va a hacer?

Figure 19. Tagging Tool

The tagged corpus was stored in a database and it is possible with a SQL query to obtain the answers to the methodological questions (1Q, 2Q and 3Q). It is noteworthy that each objective corresponds to a sentence, but it is contemplated that students type the general objective in more than one sentence. Below, the following table shows the level of Kappa agreement of the corpus, detailed in groups of 100 elements:

Table 36. Agreement level between annotators

Questions	1º 100	2º 100	3º 100	300
1Q	0.587	0.646	0.645	0.629
2Q	0.576	0.542	0.634	0.586
3Q	0.5541	0.6	0.658	0.601

The table 36 shows that the first 100 objectives (second column), obtained the level “Moderate”. The next 100 objectives (third column), 1Q and 3Q reached the “Substantial” level. The last 100 objectives, in the three methodological questions achieved “Substantial” level. The annotators achieved better performance in the last block that was tagged. Finally, the level of agreement of the 300 objectives at each level was “Substantial” for 1Q and 3Q and “Moderate” to 2Q.

5.5.1.2. Experimental results

To build the language model, the "SRI Language Modeling Toolkit[18]" was used. The "segment" command allows the identification of the segment corresponding to the question text-1Q. Below, an example of the result is shown (grammatical classes were omitted):

Table 37. Segmentation examples [19]

First example
Objective of test sample: Develop an application that supports the process of quality control, evaluating the results of each stage of the software life cycle, through the application of ISO model.
Result: **<s>** Develop an application that supports the process of quality control, evaluating the results of each stage of the software life cycle, through the application of ISO model
Text segment-1Q by annotators: Develop an application that supports the process of quality control, evaluating the results of each stage of the software life cycle
Second example
Objective of test: Design a program of auto pilot for mobile, the mobile will be able to avoid obstacles and begin evasion before crashing with them. Build a system that is able to quickly process the data sent from the sensors placed on the mobile, making the detection and evasion of obstacles more effectively. Obtain a prototype system that can be applied in real scale as a security system on highways.
Result: **<s>** Design a program of auto pilot for mobile, **<s>** the mobile will be able to avoid obstacles and begin evasion before crashing with them. **<s>** Build a system that is able to quickly process the data sent from the sensors placed on the mobile, making the detection and evasion of obstacles more effectively. **<s>** Obtain a prototype system that can be applied in real scale as a security system on highways.
Text segment-1Q by annotators: Design a program of auto pilot for mobile

One can notice that in the first evaluated objective, the segment tag **<s>** is located at the beginning of the sentence. The extension of the text segment identified by the command "segment" differs from the text identified by annotators.

[18] http://www.speech.sri.com/projects/srilm/

[19] The example has been translated to English from the original in Spanish

In the second example, the tag **<s>** appears four times, which identifies four segments. The segment identified by the annotators is very similar to the first segment identified by the segment function.

5.5.1.3. Discussion

In this experiment, it was found that the “segment” function identifies the text segments corresponding to the question 1Q. One can notice that the beginning of text-1Q segment is identified, but it is necessary to implement other strategies to identify the place where the segment ends. A strategy that could be implemented to delimit the segment would be to use the average size of the text-1Q segments of the whole corpus.

In addition, one has to obtain the value of perplexity of the text-1Q segment to verify that the segment has characteristics close to the model. To validate the segments of the text-1Q obtained from the test group, a comparison with the segments identified by the annotators will be done. A value of agreement between annotators and the identifier of answers to questions 1Q will be obtained.

As future work, we plan to identify each of the answers of the methodological questions, obtaining acceptable levels of agreement. In addition, other task will be to identify the relationships between the object and the features of the 2Q and 3Q questions, when they are found.

Chapter VI: Conclusions

6.1. Conclusions

Communicating ideas, through written language is essential to knowledge society. Many efforts to improve the students' writing at early stages of their education have been proposed. Strategies ranged from didactic support inside the classroom to the use of technology to help students improve their writing. Research studies dealing with the use of technology as an aid for writing have grown. In the study of this problem, considerable efforts were identified aiming to guide students towards better performances; for example intelligent tutoring and platforms managing the progress of students by sending personalized feedback. Moreover, automatic systems were detected which purpose is to evaluate different aspects such as lexical richness, coherence and cohesion; mainly in students' essays. In addition, a smaller amount of studies related to the internal connection of essays was pinpointed. It is noteworthy that these works address research conducted mostly in English language environments.

In this thesis, the study objects refer to theses written by university students. These documents were analyzed using different techniques of natural language processing to assist students improve their writing skills. Methods presented in this thesis were developed according to a proposed solution at four levels: Lexical Analysis, Coherence, Language Models and Methodological Questions. An important support to the development of this thesis was the Corpus consisting of theses and research proposals from the area of computing and written in Spanish.

An aspect taken into account in this thesis was the connection made between the "must be" and the NLP techniques, *i.e.* the linking of the suggestions of the authors of books on research methodology and the scope of the NLP techniques. Thus, the results obtained in this thesis are seeking to permeate the students' writing, but adhering to the guidelines on how to write a thesis.

The results obtained in each of the levels of the proposed solution showed that student's graduate level, as expected, had better performance than undergraduate level, validating in some way the different methods. NLP techniques and resources helped to solve specific problems of this thesis, however some techniques were explored in depth to adapt and achieve objectives. For some problems, the design of a specific methodology was required as the analysis of flow of concepts or the identification of weak sentences in a conclusion section.

One of research questions of this thesis refers to: How NLP techniques help to assess the main sections of a research proposal draft. Thus, after implementation of different NLP techniques, it can be concluded that the techniques allowed establishing acceptable scales to assess the different sections of the draft thesis. It was possible to connect the problem with the technique. For instance, the identification of the conceptual flow in a conclusion section was obtained using the EGrid tool, but the calculation was not immediate. However, inner workings of the EGrid technique allow to identify the conceptual flow and to connect the technique with the problem (*i.e.* identifying a conceptual flow).

In addition to the connection of the technique with the problem, the acceptable agreement level gave confidence in the results. For some cases, as lexical richness, the pilot test implemented with undergraduate students show that there was an improvement in the writing of their proposal draft.

The analysis of characteristics such as connection of general objective with the conclusions and the level of opinion in the conclusion, are internal characteristics that has not been explored before. This kind of analysis is similar, roughly speaking, to which an academic reviewer makes in their daily work. However the developed method does not seek to replace the academic advisor. In contrast, the goal is complement the academic education of students and ease the burden of teachers.

Explored solution, defined in four levels, led to the construction of a staggered solution, which was found to be adequate according to the behaviors identified in the different features analyzed in this thesis.

6.2. Further work

Each level of the proposed solution model allowed identifying different sub-problems. Besides, the feedback received from articles submitted to specialized forums broadened the overview of the application of the developed methods of this thesis. The development of this thesis has contributed to the definition of a research line whose objective is aiding students to improve writing scientific documents supported by NLP techniques.

In the future, other sections could be evaluated. For example, the Methodology, identifying the logical correspondence of the steps outlined in this section. In addition, identify the connection of the steps and techniques of the methodology with the results section.

The results that were achieved under this line in this thesis, underpin solving new challenges and exploring with deep analysis research proposals. A task to be faced is to identify the presence of arguments in drafting ideas, specifically in the section of results and conclusions.

Another future work will be the evaluation of the content, together with the assessment of the structure as developed in this thesis. For example, between the section objective and conclusions, a coverage model was explored in this thesis, however, a challenge to further work would be to assess the semantic level of coverage between those two sections.

Branching of the general problem into sub-problems helped providing solutions locally, which allowed building a model to four levels to respond to the global problem. Each level covered different aspects, first solving basic features up to dealing with complex features. In addition, the evaluated features have reached an acceptable level, allowing establish the groundwork for the methods developed in this thesis can become available to students through a computational system. The idea is that students can assess their texts with the set of methods included in the system.

Despite the fact that methods were designed for Spanish language documents, these methods can be addressed to other problems and they could be used in English language works. For instance it is possible to address other problems such as the identification of weak sentences in essays written in English by L2 students. It would however be necessary to tag a set of sentences to detect the main errors in the essays.

The guidelines for writing scientific papers are held fixed, while teaching strategies in schools have diversified and changed. The research line defined in this thesis will seek to leverage existing computing resources and adapt current characteristics of students considering the good writing guidelines.

Chapter VII: References

7. References

[1] Muñoz C. 2011. *Como elaborar y asesorar una investigación de tesis*. Pearson.

[2] Bitchener, J. and Basturkmen, H. 2006. Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes 5*, pp. 4-18.

[3] Hernández, R., Fernández, C., and Batista, M. 2010. *Metodología de la investigación*. Mc Graw Hill.

[4] Foltz, P., Kintsch, W., and Launder, T., 1998. The measurement of textual coherence using Latent Semantic Analysis. *Discourse Processes*, 25(2-3) pp. 285-307.

[5] Elsner, M., and Charniak, E. 2008. Coreference-inspired coherence modeling. Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008), pp. 41-44.

[6] Kuan-Yu, C., and Berlin, C. 2011. Relevance language modeling for speech recognition. International Conference on Acoustics, Speech, and Signal Processing. ICASSP'11, pp. 5568-5571.

[7] Montes y Gómez, M., Villaseñor, L., and López, A. 2008. Mexican experience in Spanish question answering. *Computación y Sistemas*, 12(1) pp. 40-64.

[8] Pazos, R., Gelbukh, A., González, J., Alarcón, E., Mendoza, A., and Domínguez, P. 2002. Spanish natural language interface for a relational database querying system. In 5th International Conference on Text, Speech and Dialogue. pp. 123-130.

[9] Oramas, J., and 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. p. 8.

[10] Carbonell-Olivares, M., Gil-Salomon, L., and Soler-Monreal, C. 2009. The schematic structure of Spanish PhD thesis introductions. *Spanish in Context*. 6(2) pp. 151-175.

[11] Zamora, S., and Venegas, R. (2013). Estructura y propósitos comunicativos en tesis de magíster y licenciatura. *Literatura y Linguística*. (27) pp. 201-218.

[12] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. 2010. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. 41(6) pp. 391-407.

[13] Landauer, T., and Dumais, S. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*. 104(2) pp. 211-240.

[14] Jackson, P., and Moulinier, I. 2007. Natural language processing for online applications: Text Retrieval, Extraction & Categorization. *Natural Language Processing series*. 5(1) p. 226.

[15] Kintsch, W. 2002. On the notions of theme and topic in psychological process models of text comprehension. *Converging Evidence in Language and Communication Research*. 3 pp. 157-170.

[16] Berry, M., Dumais, S., and O'Brien, G. 1995. Using linear algebra for intelligent information retrieval. *Siam Review*. 37(4) pp. 573-595.

[17] Gutiérrez, R. 2005. Análisis Semántico Latente: ¿Teoría psicológica del significado?. *Revista Signos*. 38(59) pp. 303-323.

[18] Barzilay, R., and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*. 34(1) pp. 1-34.

[19] Dagan, I., Roth, D., Sammons, M., and Massimo, F. 2013. *Recognizing textual entailment: Models and Applications*. Morgan & Claypool Publishers. p. 220.

[20] Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*. 22(2) pp. 249-254.

[21] Grobe C. 1981. Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*. 15(1) pp. 75-85.

[22] Douglas, S.R. 2010. Non-Native English speaking students at university: lexical richness and academic success. Doctoral Thesis, University of Calgary, Canada.

[23] Lemmouh, Z. 2008. The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies*. 7(3) pp. 163-180.

[24] Schwarm, S., and Ostendorf, M. 2005. Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp.523-530.

[25] Roberto, J., Martí M. and Salamó, M. 2012. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*. 48 pp. 97-104.

[26] McNamara, D., Crossley, S., and McCarthy, P. 2010. Linguistic features of writing quality. *Written Communication*. 27(1) pp. 57-86.

[27] Vilarnovo, A. 1990. Text Coherence: Internal Coherence or External Coherence?. *Estudios de Linguística Universidad de Alicante*,6 pp. 229-239.

[28] Louwerse, M. 2004. A concise model of cohesion in text and coherence in comprehension. *Revista Signos*. 37(56). pp. 41-58.

[29] Skogs, J. 2013. Subject line preferences and other factors contributing to coherence and interaction in student discussion forums. *Computers & Education*. 60(1) pp. 172-183.

[30] Medve, V., and Takac, V. 2013. The influence of cohesion and coherence on text quality: a crosslinguistic study of foreign language learners written production. *Language in Cognition and Affect*. pp. 111-131.

[31] Yannakoudakis, H., and Briscoe, T. 2012. Modeling coherence in ESOL learner texts. In: 7th Workshop on the Innovative Use of NLP for Building Educational Applications. pp. 33-43.

[32] Higgins, D., Burstein, J., Marcu, D., and Gentile, C. 2004. Evaluating multiple aspects of coherence in student essays. Human Language Technology Conference/North American chapter of the Association for Computational Linguistics. pp. 185-192.

[33] Miltsakaki, E. and Kukich, K. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*. 10(1) pp. 25-55.

[34] Dessus, P. 2009. An overview of LSA-based systems for supporting learning and teaching. Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modeling. pp. 157-164.

[35] Hernández, S. and Ferreira, 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.

[36] 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. pp. 183-188.

[37] 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. pp. 141-148.

[38] Lapata, M., Barzilay, R. 2005. Automatic evaluation of text coherence: Models and Representation. In Proceedings of International Joint Conference on Artificial Intelligence. pp. 1085-1090.

[39] Burstein, J., Tetreault, J., and Andreyev, S. 2010. Using Entity-Based Features to Model Coherence in student essays. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 681-684.

[40] O'Rourke, S., and Calvo, R. 2009. Analysing semantic flow in academic writing. Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care. pp. 173-180.

[41] Selvam, P., Natarajan, A., and Thangarajan, R. 2008. Lexicalized and statistical parsing of natural language text in Tamil using hybrid language models. *Transactions on Signal Processing*. 8(7) pp. 1362-1374.

[42] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan J. 2000. Language models for financial news recommendation. Proceedings of the Ninth International Conference on Information and knowledge Management. pp. 389-396.

[43] Wu, J., Chang, Y., Liou, H., and Chang, J. 2006. Computational analysis of move structures in academic abstracts. Proceedings of the COLING/ACL on Interactive presentation sessions pp. 41-44.

[44] Li, Y., Gorman, S. and Elhadad, N. 2010. Section classification in clinical notes using supervised hidden markov model. Proceedings of the 1st ACM International Health Informatics Symposium. pp. 744-750.

[45] Norvig, P. 1987. Inference in Text Understanding. The sixth national conference on artificial intelligence. Association for the Advancement of Artificial Intelligence. pp. 561-565.

[46] Nielsen, R., Ward, W., and Martin, J. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*. 15(4) pp. 479-501.

[47] Horbach, A., Palmer, A., and Pinkal, M. 2013. Using the text to evaluate short answers for reading comprehension exercises. Second Joint Conference on Lexical and Computational Semantics. pp. 286-298.

[48] Davis, J., and Liss, R. 2006. *Effective academic writing. 3, The essay*. Oxford University Press.

[49] Elsner, M., and Charniak, E. (2007). Brown Coherence Toolkit. Retrieved May 14, 2013, from http://cs.brown.edu/~melsner/manual.html

[50] Bethard, S., Okoye, I., Arafat, S., Hang, Martin, J., and Sumner, T. 2012. Identifying science concepts and student misconceptions in an interactive essay writing tutor. Proceedings of the 7th Workshop on Building Educational Applications Using NLP. pp. 12-21.

[51] Kilicoglu, H., and Bergler, S. 2010. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. pp. 46-53.

[52] Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9*.

[53] Eisenstein, J. and Barzilay, R. 2008. Bayesian unsupervised topic segmentation. Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 334-343.

[54] Allen, G. 1976. *The Graduate Students' Guide to Theses and Dissertations: A Practical Manual for Writing and Research,* USA. Jossey-Bass Inc Pub.

[55] Bellegarda, J. 2010. Unsupervised document clustering using multiresolution latent semantic density analysis. Workshop on Machine Learning for Signal Processing. pp. 361-366.

[56] Aiken, M., Ghosh, K., Wee, J., and Vanjani, M. 2009. An Evaluation of the Accuracy of Online Translation Systems. *Communications of the IIMA*. 9(4) pp. 67-84.

[57] Johansson, V. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Linguistics and Phonetics*. pp. 61-79.

[58] Waldvogel, D. 2014. An analysis of Spanish L2 lexical richness. *Academic Exchange Quarterly* 18(2) p. 8.

Appendix

A. Publications derived from this thesis

Publications in Journals

- González-López, Samuel; López-López, Aurelio ***Lexical Analysis of Student Research Drafts in Computing***, Computer Applications in Engineering Education, 23(4) pp. 638-644, 2015. JCR

- Submitted: González-López, Samuel; López-López, Aurelio, ***Automatic Assessment of Student Conclusion Writings: An Exploratory Study***, Computers & Education, JCR.

Publications as Book Chapter

- González López, Samuel; López-López, Aurelio ***Mining Domain Knowledge for Coherence Assessment of Students Proposal Drafts*** Alejandro Peña-Ayala (Ed.): Educational Data Mining, 524 pp. 229-255, Springer International Publishing, 2014.

- González López, Samuel; López-López, Aurelio ***Assisting Students in Writing by Examining How Their Ideas Are Connected*** Mascio, Tania Di; Gennari, Rosella; Vitorini, Pierpaolo; Vicari, Rosa; de la Prieta, Fernando (Ed.): Methodologies and Intelligent Systems for Technology Enhanced Learning, 292 pp. 9-18, Springer International Publishing, 2014.
  - ✓ We received the ***Award of Scientific Excellence***
- González López, Samuel; Bethard, Steven; López-López, Aurelio ***Identifying Weak Sentences in Student Drafts: A Tutoring System*** Mascio, Tania Di; Gennari, Rosella; Vitorini, Pierpaolo; Vicari, Rosa; de la Prieta, Fernando (Ed.): Methodologies and Intelligent Systems for Technology Enhanced Learning, 292 pp. 77-85, Springer International Publishing, 2014.

Publications in Conferences

- González-López, Samuel; López-López, Aurelio **Mining of Conclusions of Student Texts for Automatic Assessment** (Short paper), The 28th International FLAIRS2015 Conference May 18 - 20, Hollywood, Florida, USA, 2015.

- González-López, Samuel; López-López, Aurelio ***Analysis of Concept Sequencing in Student Drafts*** (Short paper), 9th European Conference on Technology Enhanced Learning, EC-TEL2014, Graz, Austria, September 16-19, Springer International Publishing, pp. 422-427, 2014

- García Gorrostieta, Jesús Miguel; González López, Samuel; López-López, Aurelio; Carrillo Ruiz, Maya ***An Intelligent Tutoring System to Evaluate and Advise on Lexical Richness in Students Writings*** (Demo), 8th European Conference, on Technology Enhanced Learning, EC-TEL2013, Paphos, Cyprus, September 17-21, Springer Berlin Heidelberg, pp. 548-551, 2013.

- González López, Samuel; López-López, Aurelio ***Supporting the Review of Student Proposal Drafts in Information Technologies*** (Full paper) Proceedings of the 13th Annual Conference on Information Technology Education (ACM SIGITE '12), Calgary , Canada, pp. 215-220, 2012.

Others Publications

- Accepted: González-López, Samuel; López-López, Aurelio ***Colección de Tesis y Propuestas de Investigación en TICs: Un Recurso para su Análisis y Estudio*** XIII Congreso Nacional de Investigación Educativa, November 2015.

- García Gorrostieta, Jesús Miguel; González López, Samuel; López-López, Aurelio ***Tutor Inteligente para Propuestas de Investigación*** Instituto Tecnológico de Aguascalientes (Ed.): Conciencia Tecnológica, Aguascalientes México, 47 pp. 43-48, 2014.

- García Gorrostieta, Jesús Miguel; González López, Samuel; López-López, Aurelio Results of a Case Study of an ***Intelligent Tutoring System for Analyzing Student Projects Presented as Research Papers***, Research in Computing Science, Center for Computing Research of IPN, 65 pp. 103-110, 2013.

- García Gorrostieta, Jesús Miguel; González López, Samuel; López-López, Aurelio ***Assessing and Advising on Lexical Richness in an Intelligent Tutoring System***, Research in Computing Science, Center for Computing Research of IPN, 56 pp. 29-36, 2012.

B. Table of Acronyms

Acronym	Description
SUM	Summarization
IR	Information Retrieval
IE	Information Extraction
QA	Question answering
AA	Answer assessment
SRA	Spanish Royal Academy

C. Glossary

Concept	Definition
Jaccard coefficient	This measure can be used to represent the similarity between two documents. The Jaccard index is defined as the intersection of the two documents divided by the size of the union of the two documents.
WordNet	This is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept.
HStO	The idea of this measure is that two lexicalized concepts are semantically close if their WordNet synsets are connected by a path that is not too long and that "does not change direction too often".
Lesk	Algorithm used to resolve the task of word sense disambiguation. The major objective of this algorithm is to count the number of words that are shared between two glosses (brief definition).
JCon	It combines a lexical taxonomy structure with corpus statistical information. The semantic distance between nodes in the semantic space constructed by the taxonomy can be better quantified with the computational evidence derived from a distributional analysis of corpus data.
Resnik	A measure of semantic similarity based on the notion of information content. Distance-based measures of concept similarity assume that the domain of documents is represented in a network.
Lin	A measure derived of Resnik similarity.
Pearson coefficient	A dimensionless index bounded between -1.0 and 1.0 which reflects the degree of linear dependence between two sets of data.
Effect size	Quantitative measure of the strength of a phenomenon.