Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 141 (2014) 470 – 483

Procedia
Social and Behavioral Sciences

WCLTA 2013

Computer-assisted revision in Spanish academic texts: Peer-assessment

Carmen López Ferrero*, Irene Renau, Rogelio Nazar, Sergi Torner

Universitat Pompeu Fabra, C/Roc Boronat, 138, 08018, Barcelona, Spain

Abstract

This paper presents a series of experiments in automatic correction of spelling and grammar errors with a statistic and corpus-driven methodology. The language of the experiments is Spanish, but the method can be easily extrapolated to other languages since we do not use language-specific resources. Our main motivation is to develop a tool that could assist university students to write academic texts, because this kind of system is practically nonexistent in the present, especially in Spanish. Our work is based on previous descriptions, which identify the most problematic phenomena in academic writing at university level. We aim to develop a tool for automatic detection and correction of some of those problematic issues at different linguistic levels such as spelling, grammar and vocabulary.

Selection and peer-review under responsibility of the Organizing Committee of WCLTA 2013.

Keywords: Academic discourse, computer-assisted revision, n-gram language models, peer-assessment, written competence evaluation;

1. Introduction and context of the study

This paper presents the main results of a project developed at Universitat Pompeu Fabra, in Barcelona, which addresses cooperative and computer-assisted revision of academic texts written by students in the first course of their University studies. The excellent domain of mother tongues (L1) in academic written discourse is a cross-curricular ability at all the University specializations: students have to write academic texts in the major part of their courses. Nevertheless, it is a general concern of teachers that, in general, students at university often show poor quality in their written expression.

* Corresponding Author: Carmen López Ferrero. Tel.: +34-935-422-310
E-mail address: carmen.lopez@upf.edu

Selection and peer-review under responsibility of the Organizing Committee of WCLTA 2013.
doi:10.1016/j.sbspro.2014.05.083

The European Higher Education Area (EHEA, http://www.ehea.info) has stated that the learning autonomy of students is a central principle. Students need to be guided with relevant resources to be able to become autonomous in their process of learning. There have been different pedagogical actions in this approach to help the student master the grammatical rules of academic written texts (on-line grammatical resources and digital written centers[†], among others), but there is not yet a computer-assisted tool to revise academic texts with accuracy. It is necessary to develop a revision method to check for the most frequent mistakes in L1 and help students be aware of the rules involved in usual grammatical problems when writing. As a consequence, it is a pedagogical tool and not only a grammar-checker what is needed in this context. Our experiments in this paper aim in that direction, and we use Spanish for them, but we are confident that the problems we are describing and the solutions we propose can also be of interest to professionals working in other languages.

We want to develop linguistic and communicative competences in Spanish for the students who enter at the University for the first time. Our proposal is to start the required tasks to build a tool for the automatic tagging of texts in order to help peer-assessment of on-line academic texts. We aim at encouraging the participation and motivation of students by involving them in their own learning process through the use of a tool that implements a tutorial of the contents of the Spanish language subject.

This paper is arranged as follows: in the next section, we describe the scope of our research. Next, section 3 presents a very brief comment on the state of the art on learner corpora and automatic error checking. Section 4 presents a detailed examination of grammatical problems that can be considered difficult for students. Section 5 presents the material and methods for our experiments and section 6 presents the results of these methods applied to the solution of some of the problems described in section 4. Finally, section 7 presents our conclusions and lines of future work.

2. Scope

The aim of our research is fundamentally educational: the main objective is to improve the correction and adequacy of academic texts written for different planned activities in the frame of a basic subject in the first year of University studies, developing a computer application to assist the revision of texts.

The following are the specific objectives of this research:

1. To improve the ability of the first year University students to write academic texts in Spanish, with the use of a computer tool to assist the revision of grammatical errors and assessment of the written texts.
2. To promote the peer-assessment of academic texts between the students.
3. To manage more dynamic and motivating assessment ways for the students, inside the process of written composition.
4. To develop revision and linguistic correction strategies which coordinate the work of students inside and outside the classroom.
5. To facilitate the evaluation, the feedback and the personal monitoring of the learning process of students.

We want to improve the training level of university students in academic writing competence. The results of our proposal could be of considerable help for students who must acquire expertise in writing academic discourse both in L1 and L2.

The computer application designed and tested in this study, which is still in a prototype stage, is based on two complementary axis: on the one hand, it is an online tool based on statistical analysis of linguistic data to correct texts automatically; on the other hand, it is a pedagogical tool to encourage autonomous learning and peer-assessment. Consequently, we consider computer applications as pedagogical tools and not only as functional tools to check errors without complementing this correction with an intervention focused on learning. The idea, thus, would be that when we arrive at the stage to develop software on the basis of these experiments, this system would not only detect and correct errors in the student's text but also would provide useful information for each case.

[†] For example, *Stilus. Corrector ortográfico, gramatical y de estilo para el español*, de Daedalus (http://www.mystilus.com/Main); *Lenguaje.com* (http://www.lenguaje.com); *Centre de Redacció* de la UPF (http://parles.upf.edu/llocs/cr).

3. State of the art: learner corpora and error analysis

The knowledge of academic writing competence has been a frequent research topic in the last few years and academic texts at the university level have been analyzed from a pedagogical point of view. Regarding only Spanish studies, several studies have described the main features of academic texts written by students as well as their main problems and have made pedagogical proposals to improve writing skills —cf. Bruck (2005), Caldera & Bermúdez (2007), Di Stefano & Pereira (2004), España (2011), Moyano (2007, 2010), Parodi (2007, 2008, 2010) and Pareira (2005), among others. Some particular aspects of academic texts have also been examined, such as several characteristic microstructures (López 2005, López & Torner 1999), syntactic maturity (López & Torner 2005, Muse *et al.* 2012), spelling (Pérez Ramos 2013) and competence in writing argumentative texts (Padilla 2004, 2009, Parodi 2000). Some other studies have focused on the linguistic characteristics of different Spanish-speaking countries: for instance, Spain (Battaner *et al.* 2001, Torner & Battaner 2005), Costa Rica (Rodino & Ross 1985, Rojas Porras 1987, Sánchez Avendaño 2005), Venezuela (Albarrán 2011), Argentina (Arnoux *et al.* 2002, Carlino 2003, 2005, 2006, 2009, 2013, Navarro 2012, Navarro & Moris 2012, Nogueira 2007) and Colombia (Vargas Franco 2005), among others. Finally, we can also find several academic writing manuals oriented to university-level students (Montolío 2000, Nogueira 2003, Vázquez 2001, 2005, among others).

Taken has a whole, these studies provide a detailed picture of university students' ability to write academic texts. However, automatic tools to assist academic writing have not yet been developed. Our work is based on these previous descriptions, specifically on the studies about *Corpus 92* (Torner & Battaner 2005), which identifies the most problematical phenomena in academic writing at university level.

The first attempts to build grammar checking tools were in the 1980s and basically for English (MacDonald 1983, Cherry & MacDonald 1983, Heidorn *et al.* 1988, Richardson & Braden-Harder 1988, Kukic 1992; cf. Leacock *et al.* 2010, for an updated review). These first grammar checkers were based on grammar rules which used complex linguistic knowledge. In spite of having been used until recently (cf. Arppe 2000, Johannessen *et al.* 2002, among others), knowledge-rich approaches do not achieve high rates of success, as noticed by Bolt (1992) and Kohut & Gorman (1995).

Some researchers adduced that the limitations of rule-based methods could be overcome by using a statistically-oriented approach. To our knowledge, the first attempt to use a statistical knowledge-poor approach was Atwell's (1987). More recent papers, such as Knight & Candler (1994) or Han *et al.* (2006), have shown that *N*-gram models can be more successful than rule-based systems. Burstein *et al.* (2004), for instance, apply an idea that has similarities to the one we present in this paper, by using *N*-grams for grammar checking. The work of Whitelaw *et al.* (2009) is another example of a method that disregards explicit linguistic knowledge for spell checking. Sjöbergh (2009) proposes a similar approach for grammar checking in Swedish, but he could not achieve a high rate of success.

Statistically-oriented approaches need a large amount of data and, as a consequence, some authors have proposed to use the web as a normative corpus (Moré *et al.* 2004, Yin *et al.* 2008, Whitelaw *et al.* 2009). In a similar way, Nazar & Renau (2012) and Atserias *et al.* (2012) proved that large corpora can be used as normative reference for statistically-based grammar checkers. The present paper is a further development of this idea. We advocate that a grammar checker based on corpus statistics methods, using very large *N*-gram corpora, could prove to be helpful. Moreover, *N*-gram approaches are language independent, so that a similar approach could be used in other languages as well.

4. Areas of study: description of the analyzed problems

In a previous project, also developed at Pompeu Fabra University during 2009 and 2011[‡], seven kinds of problems were identified in learner academic texts written in Spanish and Catalan languages: register, coherence,

[‡] *Redactant: recursos per superar mancances en la redacció de textos acadèmics*, suported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, ref. 2009MQD 00100), coordinated by Joan Costa Carreras at Pompeu Fabra University.

cohesion, vocabulary, grammar, orthography and typographic errors. In this section, a few examples of some of these frequent problems in Spanish are described to offer an idea of the difficulty of the task at hand. The selected sample can be considered representative of the kind of problems that can be usually found in student's academic texts. The problems considered here are related especially to diacritic accents in the orthographic level, to gender, prepositions and gerund in the grammatical level, and to fixed expressions in the lexical level. A description of the selected problems follows.

4.1. Orthographic level

The main problems in the orthographic level are diacritic accents. In contrast to English terminology, where the term "diacritic" is used to refer to all kinds of orthographic accents, in Spanish diacritic accent are used more narrowly to distinguish words that are otherwise homographs, the typical example being *te* pronoun and *té* ('tea'). This is not exactly the same case as *artículo* ('article') and *articuló* ('he/she articulated'), because there is here a difference in the prosodic accent in addition to the orthographic accent. From a computational perspective, however, the distinction is not crucially important because orthography does not distinguish prosody. Problems that involve the use of diacritic mark are described in what follows.

4.1.1. Confusion in spelling

4.1.1.1. porque / por que / porqué / por qué

*Esa era la causa *porque lo hizo.*
*Defiende la transparencia informativa *porqué está a favor de la democracia participativa.*

The Royal Spanish Academy (DPD 2005) details the rules of use of these items:

- *Porque* is a subordinating conjunction with two functions: a) it introduces a subordinate clause to express the cause of the principal verb action; b) it is also used as final clause followed by a subjunctive verb, with an equivalent sense to *para que*.
- *Por que* is the combination of the relative pronoun *que* preceded by the preposition *por*; the relative *que* can be substituted by the other relatives such as *el cual, la cual, los cuales, las cuales*.
- *Porqué* is a masculine noun with the meaning of 'cause or reason"; it is used preceded by article and its plural is *porqués.*
- *Por qué* is a combination of the preposition *por* and the interrogative pronoun *qué*. It is incorrect to use the article *el* before the preposition *por*: **el por qué*. It cannot be substituted by *motivo* or *razón* ('reason').

4.1.1.2. adonde / a donde / adónde

*Fue *adónde estábamos todos*
*¿*Adonde quería conducirnos?*

The word *adonde* is a relative adverb denoting place and expresses the direction of a movement; as a conjunction, it introduces relative clauses with or without antecedent; it can also be written in two words: *a donde,* as DPD (2005) specifies[§].

Adónde, written also in two words, *a dónde,* is an interrogative or exclamatory adverb with the meaning of 'a qué lugar' ('where to'). It introduces interrogative or exclamatory direct utterances or interrogative or exclamatory indirect subordinate clauses: *¿Y adónde llegaron?/ No sé adónde ir.*

[§] The Royal Academy acknowledges the use of both form in the entry for *adonde*: http://lema.rae.es/dpd/?key=adonde.

4.1.1.3. conque / con que / con qué /

*Los argumentos son muy explícitos en el texto, *con que no hay posibilidad de ambigüedad.*

The word *conque* is a deductive conjunction used to formulate a natural consequence of what have been immediately said: *Es muy nervioso, conque estad alerta.* In oral contexts, it can be used also to start an exclamation or an interrogative sentence, to express surprise or reproach to the interlocutor: *Conque esas tenemos…*

The expression *con que* is the combination of the preposition *con* followed by the relative *que: con los que, con los cuales.* It is also the combination of the preposition *con* with the conjunction que that starts the noun subordinate clauses: *Con que resulte útil, ya vale la pena dedicarle tiempo = Con resultar útil, ya vale la pena dedicarle tiempo.*

In contrast, *con qué* is the preposition *con* followed by the stress interrogative or exclamatory pronoun *qué: ¿Con qué querías trabajar si no había ningún recurso al alcance?*

4.1.2. Problems in the use of interrogative pronouns and adverbs: qué/que; cuál/cual; quién/quien; cómo/como; cuán/cuan; cuánto, cuánta, cuántos, cuántas / cuanto, cuanta, cuantos, cuantas; cuándo/cuando; dónde/donde

The written accent in the words *qué, cuál/es, quién/es, cómo, cuán, cuánto/a/os/as, cuándo* and *dónde,* as tonic units, is used when they are interrogative or exclamatory word classes, that is to say, when they are the head of sentences with an interrogative or exclamation meaning, to refer to the value of an unknown item which can be related to objects (*qué, cuál*), persons (*quién, cuál, qué*), places (*dónde*), manner (*cómo*), time (*cuándo*) or quantities (*cuán, cuánto*). Interrogative and exclamatory expressions can occur in three kinds of contexts:

a) At the beginning of a direct or indirect interrogative or exclamatory structure:

¿Cuál es el tema principal? (direct interrogation).
Es necesario explicar cómo se organiza el relato (indirect interrogation).
¡Cuánto esfuerzo! (direct exclamation).
Puede observarse cuánto cuidado hay en el tratamiento de los personajes (indirect exclamation).

It is possible that in certain contexts the interrogative and exclamatory marks are not used in independent sentences, such as in the title of a text: *Qué es la pragmática*.

b) Nominalized by a determinative: the interrogatives and exclamatory *qué, cómo, cuándo, cuánto* y *dónde* can be nominalized if they are preceded by the article *el* or by the article *un*. In those cases, they are also tonics and have a written accent: *Se require entender el qué y el cómo de su comportamiento; No hay ni un cuándo ni un dónde de los documentos encontrados*.
c) In some fixed expressions (more frequent in informal register):
   - with *qué: el qué dirán / no hay de qué / no sé qué / qué sé yo o yo qué sé / que para qué / qué va / sin qué ni para qué / sin venir a qué / un no sé qué*
   - with *cuál: a cuál más*
   - with *quién: mira quién habla o mira quién fue a hablar / no sé quién / no ser quién, no ser quiénes / quién sabe*
   - with *cuánto: no sé cuánto/a/os/as*
   - with *dónde: mira por dónde*

Moreover, the words *quién, cuál* y *cuándo* are tonic words and they are written with orthographic accent when they are used in correlative distributive clauses with an indefinite function: *quién estudia, quién trabaja, quién no hace ni una cosa ni otra.* Nevertheless, *cual* and *quien* are unstressed and they are written without orthographic accent in indefinite pronominal expressions: *cual más, cual menos; quien más, quien menos*. They also are written without orthographic accent in the indefinite pronominal expressions *cada cual* y *cada quien*, in spite of being tonic in these latter cases.

The words *que, cual/es, quien/es, como, cuan, cuanto/a/os/as, cuando* and *donde,* as unstressed units, are written without accent in the following functions:

In relative clauses, with or without expressed antecendent (*Alberti es el poeta que escribió Marinero en tierra*; *Quien conoce la materia es un especialista*)

As conjunctions: *El autor argumenta que es necesario potenciar más la creatividad*

In fixed expressions: *dar que hablar, hay que ver, como si tal cosa, cuando más/cuando menos, en cuanto a, tanto más cuanto que, tal para cual, cada quien, de cuando en cuando, de tanto en cuanto*, etc.

Finally, there are some cases where these units can be written with or without accent (ORAE 2010):

In relative clauses with an implicit indefinite antecedent with unspecific meaning:

No había donde / dónde sentarse.

Ya tengo quien / quién me acompañe.

In subordinate clauses than can be analyzed as relative ones or as indirect interrogative clauses:

Depende de cuando / cuándo sea.

In subordinate noun clauses as unstressed conjunctions (*como*) or in indirect interrogatives as tonic interrogatives (*cómo*).

Oyó como / cómo se rompían los cristales.

*P*roblems in word segmentation

- *sino* (masculine noun 'destiny', 'fate'; adversative conjunction) */ si no* (sequence formed by the conjunction *si* and the negation adverb *no*; the negation *no* is a tonic element, in front of the unstressed pronunciation of the adversative *sino*
- *sobretodo* (masculine noun 'overcoat`) / *sobre todo* (adverbial expression 'specially', 'mainly')
- *aparte*
  - adverb: 'in other place', 'separately': *Poner aparte*.
  - adjective: 'different', 'singular': *Es un caso aparte; son unos casos aparte* (always in singular).
  - noun: conversation between two or more persons to one side, separately from the rest of the presents: *Dijo en un aparte que no quería verla.*
  - preposition: *aparte de* ('leaving aside'): *Aparte de los protagonistas, otros personajes merecen ser destacados*.

  The sequence *a parte,* writen in two words, is the occasional combination of the preposition *a* and the feminine noun *parte*: *Se trata de un camino que no conduce a parte alguna* (DPD 2005); *Devolvieron los exámenes a parte de los alumnos*
- *desde / *des de* (it is always incorrect in two words).
- *entorno* (masculine noun 'environment', 'the surrounding' */ en torno* (adverbial expression: *en torno a, en torno de* 'around').

4.2. Grammatical level

4.2.1. Gender confusion of the article with some nouns

In a multilingual and international University context, there are some specific mistakes made by students of Spanish as a foreign language (L2) related with the gender of nouns, as the following examples: **la calor, *la sistema, *la análisis, *la programa, *la problema, *la tema.*

In these cases, the problem has two different origins: i) the interference of another language (*calor, anàlisi* are feminine nouns in Catalan) and ii) the incorrect interpretation of the ending of the word /a/ as a mark of feminine nouns.

Another frequent case of gender confusion of the determiner with nouns is related with the feminine nouns which start by a stressed /a/, as the following cases: **este agua, *todo el hambre, *el mismo agua.*

The compulsory use of the masculine article /el/ in this kind of feminine nouns cause, by analogy, the incorrect use of the masculine forms of demonstrative *este*, *ese, aquel*, and the determinative adjectives *todo, mucho, poco, otro*, etc. All those units have a feminine form when used with a feminine noun which starts by a stressed /a/. Moreover, the use of the article /el/ is needed only when it is near to the feminine noun which starts with tonic /a/; when there is another word between them, the concordance must be in feminine: *la misma agua*.

4.2.2. Prepositions governed by verbs

A frequent problem in academic discourse written by learners in their first year at University is the confusion in the use of prepositions required by prepositional verbs. The following are examples of phrasal verbs in Spanish with the corresponding error in the use of its prepositions: *afectar a - *afectar en; caracterizarse por - *caracterizarse porque, *caracterizarse a; corresponder a - *corresponder con; pertenecer a - *pertenecer en; preocuparse por/de - *preocuparse sobre; tratar de/sobre - *tratar Ø*.

4.2.3. Queísmo

The *queísmo* in Spanish is the incorrect elimination of a preposition (usually *de*) before the conjunction *que*, when the preposition is required by some other word in the sentence, usually a verb (cfr. DPD 2005). The following contexts are frequent cases of *queísmo* in Spanish academic texts: *acordarse (de) que; alegrarse (de) que; darse cuenta (de) que; enterarse (de) que; informarse (de) que.*

4.2.4. Gerund

The gerund in Spanish has a verbal meaning of action and an adverbial modifying function with the sense of an action that is previous or simultaneous to the action of the main verb and a subsequent action. Nevertheless, it is not correct when the subsequent meaning is not immediate to the time of the main verb of the sentence, as in the following example, where the correct form should have been *para morir en Méjico*.

*A los sesenta años (1607) emigró a América, *muriendo en Méjico, tal vez en 1614* (Seco 1986:208).

Another frequent problem in the use of gerund in Spanish language is in a specification role (not adverbial):

*Es un texto *presentando tres apartados (the correct structure should be que presenta)*

4.3. Lexical level

In the lexical level we have studied the role of fixed expressions. An example is the frequent expression *a nivel de* ('at the level of'), which cannot be used as a synonymous of 'level', 'elevation' or 'hierarchy'. For this reason, the following sentence is not correct:

**A nivel de mucosas digestivas también hay gran irritación* (DPD 2005).

The proper use would be to replace *a nivel de* in this context with the preposition *entre* or *en*. *A nivel de* is only used correctly when referring to a physical height or with the figurative sense of category or status: *Han decidido establecer relaciones diplomáticas *a nivel de embajada.*

Another example of incorrect combination is the commonly used expression in Spanish academic written text **en base a*. It must be replaced with the correct expression *con base en*:

*La petición se hizo *en base a investigaciones policiales españolas* (DPD 2005).

4.4. General remarks

In this section we have presented our preliminary investigation on a number of subjects related to typical grammatical problems that are often a challenge for students writing their academic papers. Firstly, we do not pretend to present and exhaustive list, but only a selection of those elements that we perceive as most recurrent according to our professional experience in teaching. Secondly, our idea is not to offer in this paper a solution to each of these problems because it would fall out of the scope of our research. In the present paper we are exploring the first steps in the detection and correction of grammatical errors and our first experiments are purely statistically-based. From our exposition in this section, it is evident that it will not be possible to solve all these problems just

with co-occurrence statistics because in some cases a system of grammatical rules plus lexical and morphological information will also be needed, apart from the statistical techniques. As a consequence, we see this contribution as the beginning of a longer research project, and the findings we present in this section are there to show that we do not start naively with empirical corpus experiments without a solid base of linguistic reflection.

5. Materials and methodology

5.1. Reference and test corpora for experimentation

Previous research (Atserias et al., 2012) showed that it is possible to obtain high quality results in the automatic correction of grammatical and spelling errors. In particular, these authors study errors in the accentuation of words in non-trivial cases, i.e., when both accented and non-accented forms of a particular word are listed in a lexicon. This would be the case of words such as the already mentioned example of *artículo/articuló* or other similar cases such as *ejército* ('army') and *ejercitó* ('he/she exercised'). Regular spell checkers are limited to a comparison of the words of a text with a stored vocabulary and are thus not able to correct this kind of problems, and it has been proved that it is indeed possible to tackle them by the use of just slightly more sophisticated methods. One such method is to observe which word form appears immediately before or after a target word. Using a large reference corpus, one can gather the frequency of word bigrams and determine whether a combination of two words is normal or not. In the case of one of the mentioned examples, *ella ejercitó* ('she exercised'/ 'she practiced') and *nuestro ejército* ('our army') would be examples of normal sequences of words. In contrast, combinations such as *ella ejército* or *nuestro ejercitó* are not.

The high rates of success reported by Atserias et al. (2012) are the result of an overall average of many different cases such as the above mentioned. However, a different picture emerges when one considers only a limited set of examples, those which we can classify as the most difficult cases. These are the cases of high frequency words where the previous or the following word is less helpful to detect and correct mistakes. In general, the number of different easy cases is higher than the difficult ones, but the fact that the difficult ones are very frequent makes their case interesting. High frequency words such as *cómo* ('how') or *como* ('as'/ 'such as'/ 'like') are much harder to examine because they present a greater number of word combinations (many different words can appear before or after them) and therefore they are less likely to be solved with strategies as those described in Atserias et al. (2012) or Nazar & Renau (2012). These difficult cases, such as those described in section 4 of this paper, are the ones we would like to explore now in the present research. As already said, it would be out of the scope of this work to pretend to offer a solution to all of the described cases, therefore in this study we will only focus on a short and arbitrary selection comprising the following cases: a) *cuanto / cuánto*; b) *como / cómo*; c) *cual / cuál*; d) *quien / quién*; e) *donde / dónde*; f) *porque / por que / por qué / porqué*.

One of the goals of the experiment is to determine what would be the success rate obtained using the frequency of the preceding and the following token of a target word plus the frequency of the target word itself as factors for taking a decision on whether such target word is correct or not and, in case it is not correct, to obtain proposals to correct it. With this goal in mind, we have developed an algorithm for error detection and correction and a framework to evaluate it. We have used reference and test corpus to obtain, in the first case, our model of what would be correct Spanish and, in the second, the material to evaluate how well our algorithm would perform.

As reference corpus, i.e., the corpus from which to draw reference frequencies of the combinations of words that we will deem correct, we used the Google Books Ngram corpus (Lin et al., 2012), because it is currently the largest Spanish corpus available. Because of copyright restrictions, this corpus is limited to sequences of up to five words (that is why it is called an "ngram corpus"), but for the purposes of this experiment, such context window is considered large enough. We are planning, though, to extend our reference corpus material to include other genres such as press, and with this goal we are currently working with the Spanish Wikipedia and have already downloaded a large corpus of Spanish newspapers from the web (see section 7).

As our test corpus, we used academic texts written by students, as they are the target public of our program currently in progress, together with the teachers. The corpus for experiments was the Corpus 92 (Torner & Battaner 2005), made up of around 700 tests from the *Selectividad,* the exam to enter into Spanish Universities. The corpus was sampled in the year 1992 from five Spanish cities (Barcelona, Madrid, Murcia, Oviedo, Salamanca and Sevilla),

from different domains such as history and science. This corpus has the advantage of being significantly homogeneous, as the students had the same age and level of studies. Moreover, the exams were collected in the same year and they are about the same topics. The corpus has an approximate size of 350.000 tokens.

5.2. Methodology I: using the preceding and following words as predictors

Our method to determine if a target or analyzed word is correct in a given context is based on the use of the preceding and the following word of such target word. In this phase, we use these words in the immediate context to predict if the selection of a word form in a given context is correct, based on the statistics derived from the Google Books corpus used as reference material. In the following example, the algorithm has to select the appropriate form for the context from these possibilities: *por que / porque / por qué / porqué:*

> *Y es precisamente *por qué son los primeros consumidores, que tienen que ayudar a los consumidores secundarios, los países pobres.* (Corpus 92).

The words in the immediate context are *precisamente* and *son*. In this case, the algorithm can detect the error and select the correct form (*porque*) because the both sequences *precisamente porque* and *porque son* are very frequent in the reference corpus.

5.3. Methodology II: using the prior probabilities

With the term "prior probabilities" we designate the frequency of each isolated word form to determine which is more probable with independence of its context. In isolation, this could not be enough as a method to detect errors. However, used in combination with the main method, it is a useful source of information, as evidenced from the results listed in table 1, which were obtained empirically from the reference material. The large differences in the proportion of the frequency of occurrence of each word mean that a method that always selects the most frequent form would reach high levels of precision in some cases, which means that any proposed algorithm should outperform these results in order to be considered a significant improvement.

Table 1. Examples of different proportions in the probability of occurrence of words in each set

Case	Frequency (%)
con que	93.65
conque	0.31
con qué	6.04
como	95.01
cómo	4.99
donde	92.91
dónde	7.09
que	95.82
qué	4.18

5.4. Methodology III: combining the three predictors

In subsections 5.2. and 5.3. we described three different clues that can be used as predictors for selecting a given possibility within a set of options: the frequency in corpus of the combination of a given candidate with the preceding word, the frequency of the sequence with the following word and, finally, the prior probability, which, as explained, is the relative frequency of each candidate in the reference corpus. The way we combine these three pieces of information is rather simple: in each of the three cases there is always a "winner", and given that three is an odd number, there will always be one candidate that will win the majority of the cases, unless we face the unlikely event of three different contenders each wining in one of the categories, in which case the decision is made by the one who wins in the category of prior probabilities.

For illustration, consider another example with the target word *cuál*, which in this case should be *cual*, i.e., without accent:

*Es un texto, en el *cuál se debate la decadencia de la imagen, frente al auge de la radio y los libros, como mejores medios de comunicación.* (Corpus 92).

In this example, the algorithm will compare: 1) the frequency in corpus of the sequence *el cuál* in contrast to *el cual;* 2) the frequency of the sequence *cuál se* against the frequency of *cual se* and, finally, 3) the frequency of the single word *cuál* against the word *cual*. One of the two options will win in each case, and only one of them will win in the majority of the three cases.

6. Results and evaluation

We evaluate the results of our experiment using standard measures in Computational Linguistics. Our algorithm is implemented as software that accepts text as input and the output is the same text (for this evaluation, a sentence) with the errors detected and with suggestions to correct them. In this scenario, there are four possible outcomes:

1. The input sentence is correct and the program does not flag any error.
2. The sentence contains an error which is detected and corrected by the program.
3. The input sentence is correct but the program flags it as erroneous.
4. The sentence is incorrect but the program does not find the error.

We call case 1 a "true negative" (*tn*), case 2 a "true positive" (*tp*), case 3 a "false positive" (*fp*) and case 4 a "false negative" (*fn*). Using these four values, we can calculate the figures of precision, recall, the true negative ratio, the accuracy and the F1 measure, which is the harmonic mean between precision and recall. They all provide different angles from which to assess the results. One may see precision as the probability of the algorithm of being correct when it flags an error, recall as the proportion of errors that the algorithm was able to detect, while the true negative ratio helps us to quantify the amount of noise (false positives) that the user will have to assume and accuracy will tell us how reliable the algorithm is both when it states that a sentence is correct or incorrect.

Table 2. Summary of the results of the experiments

precision	= tp / (tp + fp)
recall	= tp / (tp + fn)
true negative ratio	= tn / (tn + fp)
accuracy	= tp + tn / (tp + tn + fp + fn)
F1	= 2 . precision . recall / (precision + recall)

We conducted experiments with 100 trials (sentences) per case. These sentences from the test corpus may or not contain errors, and in fact in most cases they do not, as one could expect, because university students in general make no mistakes. The results of our evaluation are shown in Table 2. In order to measure the significance of our results in comparison to a simple method we used a baseline method based on the prior probabilities described in section 5.3. Table 3 is translated in terms of the values discussed above, resulting in the figures shown in Table 4.

Table 3. Result of the evaluation expressed in absolute numbers over 100 trials per case

	our tn	base tn	**our tp**	base tp	**our fp**	base fp	**our fn**	base fn
cuanto / cuánto	**95**	95	**5**	5	**0**	0	**0**	0
como / cómo	**88**	73	**4**	4	**3**	20	**5**	3
cual / cuál	**85**	82	**14**	14	**1**	4	**0**	0
quien / quién	**88**	76	**3**	4	**1**	14	**8**	6
donde / dónde	**92**	89	**3**	5	**1**	5	**4**	1
porque / por qué / porqué / por que	**80**	63	**15**	19	**0**	17	**5**	1

Table 4. Evaluation of the results in terms of precision, recall, true negative ratio (tnr), accuracy and F1

	our pre	base pre	**our rec**	base rec	**our tnr**	base tnr	**our acc**	base acc	**our F1**	base F1
cuanto/cuánto	**100**	100	**100**	100	**100**	100	**100**	100	**100**	100
como/cómo	**57.14**	16.16	**44.44**	57.14	**96.70**	78.49	**92**	77	**50**	25.19
cual/cuál	**93.33**	77.77	**100**	100	**98.83**	95.34	**99**	96	**96.54**	87.49
quien/quién	**75**	22.22	**27.27**	40	**98.87**	84.44	**91**	80	**40**	28.56
donde/dónde	**75**	50	**42.85**	83.33	**98.92**	94.68	**95**	94	**54.53**	62.50
porque / por qué / porqué / por que	**100**	52.77	**75**	95	**100**	78.75	**95**	82	**85.71**	67.85

As we can see in tables 2 and 3, the algorithm performs significantly better than the baseline, except in the first case (*cuanto / cuánto*) because with both strategies we can detect and correct the only five errors that were among the 100 sentences. But in the rest of the cases, we have a good balance between precision and recall or, as we can say, silence and noise. Our algorithm produces significantly less noise (false positives) than the baseline.

We can also appreciate that there are important differences in performance among cases. For instance, the precision in the case of the pair *como / cómo* is significantly worse than the rest of the cases. Still, in this case also the precision in much higher than the baseline. In general, the comparatively low rate of false positives and the fact that the values of true negative ratio and accuracy remain high mean that in a real life scenario our system would not have a very serious impact of false alarms, which is one of the greater risks of a system of these characteristics. Of course, a system that too often flags errors in sentences that are correct would surely be deemed as not useful by real users. The risk of letting errors pass unnoticed is, however, a factor that we will have to address, as there are in some cases room for improvement in the figures of recall. This is where a supplementary system of rules would be most useful.

7. Conclusions and future work

We have presented an investigation into some of the most typical classes of grammatical errors presented by students in Spanish academic texts and our first experiments towards the development of software for grammatical error detection and correction. Although still very preliminary, the results of our experiments already show that our method is valid and the line of research fruitful, deserving thus to be further developed.

We do not claim that a method based purely on simple corpus statistic will be sufficient for full fleshed software on grammatical error detection and correction. On the contrary, we have stated that this strategy will have to be integrated with a system of linguistic rules including grammar, lexical databases and a morphological analyzer. In this attempt, however, it was important to separate the results that can be obtained with a purely statistically-oriented algorithm, because this is the simplest approach and the one that can be easily extrapolated to experiments with other languages.

Overall, the results achieved so far in this research are already a valid contribution to our initial goal: to develop a tool that would enable students to auto- and co-evaluate their academic papers in the process of peer-assessment. With that goal in mind, and as already mentioned in the introduction, ours will be a truly pedagogical tool in that, apart from being able to detect errors and propose corrections, it will also offer grammatical information related to the detected mistake.

Our research program now continues in several directions. First of all, we are now replicating our experiments using more reference corpora such as the Spanish Wikipedia and a 5 billion word corpus of Spanish press articles downloaded from the web, in order to quantify the improvement that can be achieved with more reference material. Once we have obtained estimates based on the different statistics, we will then determine if our system can be improved on the basis of a set of grammar rules to systematize the correct usage of the language. We also plan to extend our experiments with a large scale evaluation including more cases of each class and more classes of cases comparing the results with what is currently being offered by other grammar checkers such as those included in Microsoft Word and Google Docs.

Acknowledgements

The paper received funding from project 20 PlaCQUID 2012-2013 1, from Universitat Pompeu Fabra.

References

Albarrán Santiago, M. (2011). La calidad de la prosa escrita producida por el alumnado de bachillerato. *Revista de Evaluación en Investigación,* 6(2), 103-114.

Arnoux, E., Di Stefano, M., & Pereira, C. (2002). La lectura y la escritura en la universidad. Buenos Aires: Eudeba.

Arppe, A. (2000). Developing a grammar checker for Swedish. *Proceedings of the Twelfth Nordic Conference in Computational Linguistics.* Trondheim, Norway, 5–77.

Atserias, J., Fuentes, M., Nazar, R., & Renau, I. (2012). Spell checking in Spanish: the case of diacritic accents. *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA).

Atwell, E. S. (1987). How to detect grammatical errors in a text without parsing it. *Proceedings of the Third Conference of the European Association for Computational Linguistics,* Copenhagen, Denmark, 38–45.

Battaner, M. P., Atienza, E., López, C., & Pujol, M. (2001). Aprender y enseñar. La redacción de exámenes. Madrid: Machado Libros.

Bolt, Ph. (1992). An evaluation of grammar-checking programs as self-help learning aids for learners of English as a foreign language. *Computer Assisted Language Learning*, 5(1), 49–91.

Bruck, C. (Ed.). (2005). *Español con fines académicos: de la comprensión a la producción de textos.* Madrid: Edinumen.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: the criterion writing service. *AI Magazine,* 25(3), 27–36.

Caldera, R., & Bermúdez, A. (2007). Alfabetización académica: comprensión y producción de textos. *Educere,* 11(37), 247-255.
Carlino, P. (2003). Alfabetización académica: Un cambio necesario, algunas alternativas posibles. *Educere,* 6(20), 409-420.
Carlino, P. (2005). *Escribir, leer y aprender en la universidad. Una introducción a la alfabetización académica,* Buenos Aires: Fondo de Cultura Económica.
Carlino, P. (2006). Concepciones y formas de enseñar escritura académica. Un estudio contrastivo. *Signo y Seña,* 16, 71-117.
Carlino, P. (2009). Prácticas y representaciones de la escritura en la universidad: los casos de Australia, Canadá, EE. UU. y Argentina. *Cuaderno de Pedagogía,* 6, 6-17.
Carlino, P. (2013). Alfabetización académica. Diez años después. *Revista mexicana de investigación educativa,* 18(57), 355-381.
Cherry, L., & McDonald, N. (1983). The writers workbench software. *Byte,* 241–248.
Di Stefano, M., & Pereira, C. (2004). La enseñanza de la lectura y la escritura en el nivel superior: procesos, prácticas y representaciones sociales. In P. Carlino (Ed.), *Leer y escribir en la universidad.* Buenos Aires: Asociación Internacional de Lectura Lectura y Vida, 23-39.
España Palop, E. (2011). Enseñanza de la escritura en el ámbito universitario: situación actual y perspectivas. *Normas. Revista de estudios lingüísticos hispánicos,* 1, 37-51.
Han, N. R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering,* 12(2), 115–129.
Heidorn, G., Jensen, K., Miller, L., Byrd, R., & Chodorow, M. (1982). The epistle text-critiquing system. *IBM Systems Journal,* 21, 305–326.
Johannessen, J. B., Hagen, K., & Lane, P. (2002). The Performance of a Grammar Checker with Deviant Language Input. *Proceedings of the 19th International Conference on Computational Linguistics.* Taipei, Taiwan, 1–8.
Knight, K., Chandler, I. (1994). Automated Postediting of Documents. *Proceedings of National Conference on Artificial Intelligence,* Seattle, USA, 779–784.
Kohut, G. F., & Gorman, K. J. (1995). The effectiveness of leading grammar/style software packages in analyzing business students' writing. *Journal of Business and Technical Communication,* 9, 341–361.
Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys,* 24, 377–439.
Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. USA: Morgal and Claypool.
Leacock, C., Chodorowm, M., Gamon, M., & Tetreault, J. (2010), *Automated grammatical error detection for language learners, Synthesis lectures on human language technologies*, 2010, Vol. 3, No. 1, 1-134.
Lin, Y, Michel, J.B., Lieberman Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Vol. 2: Demo Papers (ACL'12).
López Ferrero, C. (2005). Funciones retóricas en la comunicación académica: formas léxicas de modalidad y evidencialidad. *Signo y Seña* 14, 115-139.
López, C., & Torner, S. (1999). Disponibilidad léxica y ponderación en el discurso académico. El uso de los adjetivos en el Corpus 92. *REALE. Revista de Estudios de Adquisición de la Lengua Española,* 11, 23-45.
López, C., & Torner, S. (2005) Rasgos sintáctico-discursivos en el Corpus PAAU 1992: aproximación cuantitativa. *Lingüística Española Actual,* 27(1), 219-248.
MacDonald, N. H. (1983). The UNIX Writer's Workbench Software: rationale and design. *Bell System Technical Journal,* 62, 1891–1908.
Montolío, E. (Coord.) (2000). *Manual práctico de escritura académica.* Bacelona: Ariel.
Moré, J., Climent, S., & Oliver, A. (2004). A grammar and style checker based on Internet searches. *Proceedings of LREC 2004,* Lisbon, Portugal.
Moyano, E. (2007). Enseñanza de habilidades discursivas en español en contexto pre-universitario: una aproximación desde la LSF. *Signos,* 40(65), 573-608.
Moyano, E. I. (2010). Escritura académica a lo largo de la carrera: un programa institucional. *Signos,* 43(74), 465-488.
Muse, C. E. M., Delicia, D. D., Fernández, M. V., & Porporato, G. (2012). Madurez sintáctica en estudiantes universitarios: un estudio comparativo sobre la producción del discurso académico oral y escrito. In I. V. Bosio, *et al.* (coord.) *Discurso especializado: estudios teóricos y aplicados*. Mendoza: Universidad Nacional de Cuyo, 209-222.
Navarro, F. (2012). Alfabetización avanzada en la Argentina. Puntos de contacto con la enseñanza-aprendizaje de español académico como L2. *Revista Nebrija de Lingüística Aplicada,* 12(6), 49-83.
Navarro, F., & Moris, J. P. (2012). Estudio contrastivo de monografías escritas en las carreras de Educación, Filosofía, Historia y Letras. In I. V. Bosio, *et al.* (coord.). *Discurso especializado: estudios teóricos y aplicados.* Mendoza: Universidad Nacional de Cuyo, 151-160.

Nazar, R., & Renau, I. (2012). Google Books N-gram Corpus used as a grammar checker. *Proceedings of EACL 2012: Second Workshop on Computational Linguistics and Writing (CL&W 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering.* April 23, 2012, Avignon, France.

Nogueira, S. (Coord.) (2007). *La lectura y la escritura en el inicio de los estudios superiores. Prácticas de taller sobre discurso académico, político y parlamentario.* Buenos Aires: Biblos.

Nogueira, S. (Ed.). (2003). *Manual de lectura y escritura universitarias.* Buenos Aires: Biblos.

Padilla, C. (2004). Producción de discursos argumentativos en estudiantes universitarios. *Actas del V Congreso de Lingüística General.* Madrid: Arco/Libros, 2193-2202.

Padilla, C. (2009). Argumentación académica: la escritura de ponencias en el marco de una asignatura universitaria. In Núñez, *et al.* (Eds.), *Actas del XI Congreso de la Sociedad Argentina de Lingüística.* Santa Fe (Argentina): Universidad Nacional del Litoral.

Parodi, G. (2000). La evaluación de la producción de textos escritos argumentativos: una alternativa cognitivo/discursiva. *Signos,* 33(47), 151-166.

Parodi, G. (2007). El discurso especializado escrito en el ámbito universitario y profesional: constitución de un corpus de estudio. *Signos,* 40(63), 147-168.

Parodi, G. (2008). Géneros académicos y géneros profesionales: accesos discursivos para saber y hacer. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (Ed.) (2010). Alfabetización académica y profesional en el siglo XXI: leer y escribir desde las disciplinas. Santiago de Chile: Academia Chilena de la Lengua / Planeta.

Pereira, M. C. (Ed.) (2005). La comunicación escrita en el inicio de los estudios superiores. Los Polvorines: UNGS.

Pérez Ramos, L. (2013). Análisis del cuidado de la ortografía en los textos académicos de los alumnos de primero de bachillerato. *Memoria final del Máster en Profesorado de Educación Secundaria.* Universidad de Almería.

Richardson, S., Braden-Harder, L. (1988). The experience of developing a large-scale natural language text processing system: CRITIQUE. *Proceedings of the Second Conference on Applied Natural Language Processing (ANLC'88).* ACL, Stroudsburg, PA, USA, 195–202.

Rodino, A. M., & Ross, R. (1985). Problemas de expresión escrita del estudiante universitario costarricense. Un estudio de lingüística aplicada. San José: EUNED.

Rojas Porras, M. (1987). Análisis del nivel discursivo: Registros escritos de undécimo año. *Revista Educación,* 11 (2), 15-22.

Sánchez Avendaño, C. (2005). Los problemas de redacción de los estudiantes costarricences: una propuesta de revisión desde la lingüística del texto. *Filología y Lingüística,* XXXI (1), 267-295.

Sjöbergh, J. (2009). *The Internet as a normative corpus: grammar checking with a search engine.* Techical Report, Dept. of Theoretical Computer Science, Kungliga Tekniska Högskolan.

Torner, S., & Battaner, M. P. (eds.) (2005). El Corpus PAUU 1992: estudios descriptivos, textos y vocabulario. Barcelona: Institut Universitari de Lingüística Aplicada (Sèrie Monografies, 9).

Vargas Franco, A. (2005). Escribir en la universidad: reflexiones sobre el proceso de composición escrita de textos académicos. Cali: Universidad del Valle.

Vázquez, G. (Coord.) (2001). ADIEU. Discurso académico en la Unión Europea: Guía didáctica del discurso académico, El discurso académico oral y Actividades para la escritura académica. Madrid: Edinumen.

Vázquez, G. (Coord.) (2005). Español con fines académicos. Madrid: Edinumen.

Whitelaw, C., Hutchinson, B., Chung, G. Y., & Ellis, G. (2009). Using the web for language independent spell checking and autocorrection. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 890–899.

Yin, X., Gao, & Dolan, W. B. (2008). A web-based English proofing system for English as a second language users. *Proceedings of the 3rd International Joint Conference on Natural Language Processing,* Hyde