International Journal of Linguistics and Communication
June 2014, Vol. 2, No. 2, pp. 13-42
ISSN: 2372-479X (Print) 2372-4803 (Online)

Published by American Research Institute for Policy Development

A Corpus-Based Writing Aid for Spanish Language Authors: The Scientific Abstract Generator Prototype[1]

Belén López Arroyo[2] and Roda P. Roberts[3]

Abstract

Abstracts are a secondary genre based on the Research Paper (RP) that have often been analyzed in English for insights into rhetorical structure and information distribution. However, this wealth of descriptive research has not produced particularly useful results for scientists who are not native speakers of English nor has it been "directly amenable to applied endeavours". The aim of this paper is to describe the methodology and the tools devised by the ACTRES research group to bridge the transition between the descriptive and the procedural approach. The findings obtained will feed into a writing application for Spanish-speaking scientists who need to report their work in English to the global research community: it is called the *Scientific_Abstract_Generator*. A custom-made comparable corpus, BioABSTRACTS_C-ACTRES, has been compiled and analyzed for rhetorical and lexico-grammatical features of this genre in both English and Spanish. Then, cross-linguistic similarities and differences relevant for our intended users have been identified and will be used to build a writing prototype available as a useful and usable computer interface.

Keywords: scientific abstracts, genre studies, corpus-based studies, contrastive studies, text generator

[1] Research for this article has been undertaken as part of the ACTRES program, partly funded by the Ministry of Education and ERDF [FFI 2009-08548] and the regional government of Castile-Leon [LE–025A09].
[2] Universidad de Valladolid (Spain). ACTRES Research Group, Address: Dpto. Filología Inglesa. Facultad de Filosofía y Letras, Plaza del Campus s/n. Valladolid, 47011 Spain. E-mail: belenl@lia.uva.es, Phone number + 34 629125260, Fax number: + 34 983423747
[3] University of Ottawa (Canada). ACTRES Research Group, Bilingual Canadian Dictionary Project, University of Ottawa-ACTRES, 40 Stewart, Ottawa, K1N 6N5 Canada. E-mail: roberts@uottawa.ca

1. Introduction

In this era of globalization, where English is the *lingua franca* for scientific dissemination, scientists whose first language is not English and may have problems in cross-cultural written communication demand applications or tools that help them do their job properly rather than theoretical or descriptive studies. In other words, what "users require, and indeed desire, are ready-to-use aids" (Rabadán 2008, 309); that is to say, the user is demanding *usefulness*, understood as "the extent to which tools (technological, conceptual or otherwise) satisfy the actual needs of a user" (Rabadán, 2008: 106; Landauer, 1995: 4) and usability or "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241; Quesenbery, 2001; Kreitzberg and Little, 2009).

ACTRES group (http://actres.unileon.es), aware of this situation, has decided to focus their research on attempting to fill this niche which has been neglected in other research programs; ACTRES research tries to cater for a pressing need of non-linguist users: namely, to have available Spanish-English bilingual written communication aids designed for particular professional groups whose first (or dominant) language is Spanish. Thus, after descriptive, contrastive and evaluative stages not only of relevant genres for the discourse community, but also of the user's needs, ACTRES group decided to design some tools that helped researchers disseminate their knowledge. These aids are envisaged as user-friendly computer applications, based on templates and termed 'generators' that will enable non-linguist users to make correct decisions on the basis of corpus-validated contrastive research. They will consist of an interface giving access to (i) text-linguistic guidelines, and (ii) phraseological information.

The prospective users for this application have been identified as Spanish-speaking experts in a particular specialized field whose mastery of English (in general and for their particular purposes) can be ranked as B1/ B1+ on the scale of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2000). These prospective users generally report the following claims when writing in English:

- They always start from Spanish and try to fill in the lexical slots with equivalent resources, often obtained from free online machine translation systems.
- Their main problem is not expert terminology, but 'stringing words together'.

- They are not competent to assess whether their abstracts are linguistically correct or textually acceptable to native Anglophone readers.
- Their poor performance in written English limits their professional advancement, because they miss out on publication and promotion opportunities.
- Lack of funding and/or time constraints do not always allow for recourse to language services.

Our task is then to find solutions to the above problems, namely, to design a writing aid that (i) is source-language based, i.e. that starts from Spanish, (ii) focuses on phrases rather than terms, and (iii) ensures that the resulting text is both correct and acceptable to Anglophone readers. In addition, this writing aid has to be easy to use, and adapted to the capabilities of its end-users. In short, usefulness and usability must inform the decisions made throughout the research and development process.

This goal raises questions concerning text and user profiling, and issues of applicability. It also involves deciding which instruments, technical or otherwise, are appropriate and whether to use off-the-shelf tools (if available) or custom-made ones. A third issue to be addressed is the methodological procedure to be followed when analyzing the empirical data. And finally, the technological tool that constitutes the template of our writing aid prototype has to be accounted for.

This paper sets out to address all of the above issues. Empirical work will concentrate on abstracts, a genre chosen because of its role in the dissemination of research.

2. Abstracts as Metatexts

Communication between scientists has its own rhetoric, which means that scientists in any given field use text-internal and text-external features that the ordinary speaker of a language will not be familiar with, because specific domain meanings have been assigned to them. These features share a basic meaning and/or function across languages, but they differ in the way information is distributed and in the resources used. Rabadán (2002: 39) maintains that languages show preference for particular text-linguistic strategies, observed in the distribution and frequency of certain structural, semantic and pragmatic rhetorical patterns.

Consequently, knowing and understanding these patterns of linguistic behaviour is essential for those experts whose first language is not English, if they want their research and findings to be successfully reported and received.

According to Swales (1990: 25-26) a discourse community is a group of users characterized by six defining features: (i) "an agreed set of common public goals", (ii) "mechanisms of intercommunication among its members", (iii) "mechanisms to provide information and feedback", (iv) "one or more genres in the communicative furtherance of its aims", (v) "specific lexis", and (vi) a reasonable number of members with "a suitable degree of relevant content and discoursal expertise". The global scientific community, whose *lingua franca* is English, complies with all these conditions, and being able to produce an abstract is part of the "discoursal expertise" that can be expected from members of the scientific community.

Swales' definition of genre as a group of texts which share a set of communicative purposes and are recognized as having legitimacy within a discourse community (1990: 58) also applies to abstracts, which are generally considered as a subsidiary genre with a function that is relevant for the scientific discourse community.

Abstracts are based on a primary text, the Research Paper (RP), and normally function as metatexts (Nord, 1997: 54); that is, abstracts accurately and briefly show the contents of the RP from which they derive. Abstracts may be classified on the basis of content, structure and authorship as well as purpose or function. According to their function, two standard types of abstracts *descriptive* and *informative*, can be distinguished, Russell (1988: 4).

Descriptive abstracts help "readers understand the general nature and scope of the RP but they do not go into a detailed step-by-step account of the process involved" (Lorés, 2003: 74), whereas informative abstracts contain more detailed information on purpose, scope, methods, results and conclusions or recommendations, and therefore they have to be divided into the same sections as RPs, following what is known as "the IMRD pattern."[4]

[4] Scientific journal generally requires articles to follow the IMRD pattern, and abstracts must follow the same pattern. IMRD stands for Introduction, Methods, Results and Discussion. Each section has its own rhetorical structure, and readers of RPs and abstracts expect writers to adopt this structure.

Sometimes, informative abstracts are the only part of a piece of writing that readers actually read. A number of publications (i.e. *Translation Studies Abstracts Online*) publish only abstracts. They do so as a source of quick information and orientation. Therefore, producing well-written abstracts in English becomes increasingly important for directing readers to articles of potential interest and value (López Arroyo, 2007: 8). Not being able to do so means that non-Anglophone scientists are less likely to draw readers to their RPs. Helping them calls for new writing aids and applications, and in order to meet this need, we first need to profile the typical user of such an aid or application.

3. Methodology: Tools

As is customary in ACTRES research projects, this project uses conceptual, technical and evaluative tools (Rabadán, 2008a) as its methodological framework.

3.1 Conceptual Tool: Rhetorical Labels as Tertium Comparationis

As has been reported in several functional contrastive analysis approaches (Chesterman, 1998; 2007; Rabadán, 2002; 2007; 2008; 2008a among others), any cross-linguistic study has to establish beforehand some criteria for the comparison; this is generally referred to as the *tertium comparationis* (Krzeszowski, 1990) and in the present study, it consists of a number of cross-linguistic rhetorical labels whose role is to mark the internal divisions in the genre as well as to identify both intra-linguistically and cross-linguistically all the rhetorical units constituting abstracts. Because the goal is an applied one, these labels do not need to follow any particular model of discourse analysis nor be constrained by the methodological limitations of a given linguistic theory. Instead, any sources that may contribute to the discriminatory power of the analysis may be drawn upon and new labels proposed whenever useful. In this particular case, it seems appropriate to adopt the hierarchical rhetorical labels of moves and steps first put forward by Swales (1990) and subsequently adopted by Nwogu (1997), Bhatia (2004) and Biber et al. (2007) among others.

Biber et al. (2007: 24) consider that the general organization patterns of texts are best described as a series of units called moves, a move being a meaningful element realized by linguistic (lexico-grammatical) means which fulfils a communicative function (Biber *et al.*, 2007: 23).

A move not only achieves partial objectives, but also contributes to the fulfilment of the overall communicative purpose of each genre (Henry and Roseberry, 2001a: 95); it may contain lower-level functional-cognitive constituents, named steps according to Swales (1990; 2004) or strategies according to Bhatia (1993), that, either together or in a particular combination, contribute to achieve the communicative purpose of the move to which they belong (Biber *et al.*, 2007: 24) and compose the rhetorical structures.

Thus, our *tertium comparationis* is a set of rhetorical labels for all the elements constituting the abstract, that is to say for moves, steps and sub-steps, if any. This is then a top-down analysis that needs a coding protocol; in other words, once the rhetorical purpose of the genre has been understood, move and step categories need to be established and a set of labels for those categories determined. The labels play a diagnostic role in the contrastive procedure by helping to locate the rhetorical elements that are different in the two languages. In the case of scientific informative abstracts, their conventional sections (Introduction, Materials and Methods, Results and Conclusion) may consist of up to three different moves. For example, the section of Conclusion may include two moves, Specific research outcome and Research conclusions, and the move of Research Conclusion can be divided into two steps, Implications and Further research, which may in turn be divisible into smaller sub-steps.

Segmenting abstracts into moves and steps has to be done manually because it requires human judgment (Biber *et al.*, 2007: 35). For this reason, we have developed an assessment process, which uses informants both to achieve minimal objectivity and to ensure consistency in the analysis. "see Section 3.3"

3.2 Technical Tools: Building the BioABSTRACTS_C-ACTRES Corpus

Our primary technical data gathering tool is a comparable corpus, which will supply empirical data about correct usage in the two languages (Rabadán, 2008a: 311). Existing general or specialized corpora are likely to provide data for features typical of broad genres, although they rarely document every sub-genre (Aston, 1999: 292). As it was impossible to find existing corpora containing abstracts in both English and in Spanish, we had to compile an *ad hoc* corpus. By *ad hoc* corpus or customized corpus (Austermühl, 2001: 128) we understand a corpus compiled "on the fly" (Aston, 1999: 290) to investigate a specific problem or to carry out a specific analysis.

Comparable corpora provide "real language for real cross-linguistic communication problems" (Rabadán, 2008: 107), which makes them a fundamental tool to reveal cross-linguistic differences in problem areas (Rabadán, 2007: 239), in this case rhetorical and phraseological differences in the construction of scientific abstracts.

The size of the corpus deserves mention, since the aim of the study and the methodology determine the type and size of corpus to be used. Given that small corpora have proven useful for LSP and pedagogical purposes (Flowerdew, 2005: 329), and taking into consideration the semi-manual methodology designed and used in the present study, it was decided to make the BioABSTRACTS_C-ACTRES corpus between 10,000 and 15,000 words per language. However, both the number of texts (50 in each language) and the word count (14,484 words in English and 15,113 in Spanish) exceed Biber's (1993: 254) proposals of "at least twenty texts per register" and between 2,000 and 5,000 words, and meet Bowker and Pearson's (2002: 48) more ambitious requirements of "anywhere from about ten thousand to several hundreds of thousands of words".

Other prevalent criteria for the compilation of corpora are representativeness and availability. The abstracts were chosen from highly rated databases so as to ensure a representative sample of the language of the discourse community (Nwogu, 1997: 121). Availability refers to the ease with which the abstracts can be obtained.

As for representativeness and availability, we started our search on the Internet in broad scientific community databases. However, in a second stage, we restricted our search to more focused search engines such as Medscape, which selects abstracts and research papers depending on their scientific validity, importance, originality and contribution to the scientific community, i.e., to medicine. For our English subcorpus, each journal had to meet at least one of the following criteria, over and above the Medscape selection criteria:

- be highly ranked according to the expert opinion of pre-eminent clinicians and researchers
- be one of the nine English-language international general medical journals whose full-time editors are members of the International Committee of Medical Journal Editors

- be included in the 1994 internal JAMA (*Journal of American Medical Association*) journal list
- have a journal impact factor greater than 2 as ranked by the Institute for Scientific Information's Journal Citation Reports
- boast high readership scores as determined by PERC (Pharmaceutical and Health Care-related Promotion Research).

However, all these criteria could not be applied when compiling the Spanish subcorpus. For example, international impact could not be used because ISI does not include any Spanish medical journals in its ranking. Nevertheless, Medscape covers two Spanish journals and so we took the abstracts derived from those two journals: *Revista Española de Cardiología,* the official publication of the Sociedad Española de Cardiología*,* and *Medicina Clínica***,** an Elsevier publication.

A further question is annotation. Once the texts were selected, downloaded and filed, they were stripped down to plain text (.txt) format for processing with corpus software. The files were tagged at two levels; the first level provides identification information for the texts, such as journal or other origin and date of publication; and the second level provides a set of "move tags" (Tribble 2001: 389). These "move tags" were named rhetorical labels in the present study in order to differentiate them clearly from the tags of the first level.

In order to clarify the former, here is an example tag: [001AbIJC110000BioSciEn], where 001 represents the sample's order number in the corpus; Ab stands for the genre, namely abstract; IJC identifies the origin of the example, in this case the *International Journal of Cardiology*; 110000 stands for the date, 2011; BioSci marks the field, Bioscience; and En or Es identifies the language, English or Spanish respectively. The rhetorical tags will be established when the methodology procedure is carried out. "see Section 4".

Thus, our corpus can be described as comparable, bilingual, synchronic and annotated: comparable because all the source abstracts are on medical topics; bilingual because they are original writings in English or Spanish; synchronic because they were all published within the last decade; and annotated, because we tagged the identification information as well as marked the rhetorical structure of every abstract in order to encode similarities and differences in their construction.

3.4 Evaluation Tool: Informants

Our evaluative tool consists of two sets of informants, one for language and the other for usability.

Our language assessors play an important role at two different stages, namely, when analyzing for meaning and when assessing typicality, understood in the present study as 'the linguistic options regularly chosen by the speakers of a language at a particular time and in a particular sociocultural situation' (Labrador, 2003: 41). Their sociolinguistic profile has been described elsewhere (Rabadán, 2008a) and what follows is merely a rough characterization: 'university educated', 'middle class', age range '25-55 years'. They are all native speakers of either English or Spanish and have some training in linguistic analysis, although only one in each language group is a professional linguist. In each subgroup there is at least one person who is not bilingual. The rest can communicate in both languages at varying levels of proficiency.

The second set of informants is representative of our intended end-users: engineers, health workers and academics in the biosciences, whose proficiency level in English reading and understanding is at least B1+ CEFR (Council of Europe 2000), and who can be considered as belonging to the same societal segment as the language informants. Their task in our research protocol is to help identify usability issues, linguistic or otherwise, at every stage of the process. So far they have not required extra training in order to perform their function.

4. Procedure and Experimentation

4.1 Qualitative Data: The Rhetorical Moves

In the present study, we have adopted Nwogu's (1997) schematic structure of information for medical RPs as well as Harris & Hofmann's (1973) abstract templates as a starting point, since informative abstracts derive from RPs. However, previous cross-linguistic studies on abstracts (Méndez and López, 2003; López & Méndez, 2007) have shown that abstracts do not include all the moves and steps of the RPs, because of their concise nature. For example, whereas RPs include 3 or 4 moves for each section, RP abstracts include 2 or 3 optional moves. Nwogu's proposal has therefore to be refined and adapted to the empirical data obtained from the corpus.

Texts were then manually labelled and to minimize any bias due to the coding by a single individual, the labelling was done by two researchers at different times, using language informants "see Section 1" whenever necessary. Finally, we reconciled the differences between the researchers' and the language informants' preferred labelling in order to achieve what Biber et al. (2007: 35) call "inter-rater agreement: 82.1%"

4.2 Quantitative Data: a Hierarchy of Prototypical Moves

This qualitative procedure will be followed by the quantitative corpus analysis, as recommended by Flowerdew (1998), Biber *et al.* (1998) and Upton and Connor (2001). These authors recommend the combination of quantitative and qualitative analysis when working with a corpus, because this approach provides much of the qualitative detail that is of value to genre analysis while adding to it the reliability that is best assured by the quantitative power of computer analysis (Upton, 2002: 66). The purpose of the latter is to distinguish the most recurrent moves from the secondary ones by the frequency of occurrence of each rhetorical move in the abstracts. The most frequently recurrent moves, which range between 40% and 100% frequency, are considered conventional (Biber et a*l.*, 1997: 24) or compulsory (Suter, 1993: 119). This category includes Suter's compulsory high-priority and medium-priority moves and steps as listed below. The moves occurring less frequently (<40%) are deemed low priority and occasional and are called optional. Overall, the moves comprise

- compulsory moves and steps (C): appearing in between 80% and 100% of their section or move;
- high priority moves and steps (HP): between 60% and 80% of their section or move;
- medium priority moves and steps (MP): between 40% and 60% of their section or move;
- low priority moves and steps (LP): between 20% and 40% of their section or move;
- occasional moves and steps: appearing in less than 20% of their section or move.

The moves and steps with the highest occurrences, i.e. C, HP and MP, are considered the most prototypical. Therefore only moves with more than 40% frequency will be included in one of our tools the *Scientific_Abstract_Generator*.

5. Contrasting Texts: Stages

The methodology has been applied to our comparable corpus following Kreszowski's approach (1990: 35) of description, juxtaposition and contrast. The first stage involves a intra-linguistic description in order to find evidence of typicality and of the distribution of the resources available to express a given meaning in each language separately. Hence, the description is concerned with the "quantitative and qualitative analysis of empirical data retrieved from the corpus" (2008: 112).

The juxtaposition of the data obtained in the descriptive phase highlights the differences in the construction of the abstracts, i.e. interlinguistic divergences, and supplies "information about whether a resource is more central (or prototypical) in one language than in the other" (2008: 113). Finally, comparing and contrasting the results establishes a degree of overlap between uses, and aims at developing "an inventory of descriptively correct possibilities" (Rabadán, 2007: 249). The analysis proceeds from the identification of rhetorical moves and steps to lexico-grammatical choices in all the phases.

5.1 Description

The description phase starts by the pilot segmentation of the abstracts of each subcorpus into moves and steps so as to identify their prototypical rhetorical structure for the abstracts and their distribution in each language. The aim of this step is to obtain information on the relative frequency of particular moves and steps and the prototypical and alternative patterns of move type usage in each language. This information is essential for developing a prototype.

Once this step is completed, the samples are labelled and uploaded to the ACTRES Tagger "cf. Figure 1", a computerized tool which provides the researcher with tags specific to a genre (abstracts, minutes...) and a field (biosciences, technology, economics...). Once the tags are added, the tagged abstracts are exported to an .xml file for further processing.

Figure 1: Tagger Tool

Table 1 below shows the results of the quantitative and qualitative analysis (structure and distribution) of the abstracts. The classification of moves and steps as high, medium, compulsory or low priority has also been included. Prototypical information is shown in bold. "cf. Table 1"

Table 1: Rhetorical Elements of English and Spanish Scientific Abstracts and their Most Prototypical Rhetorical Structure (For a key to the Abbreviations see Appendix 1)

INTR Moves	English Subcorpus distribution	Spanish Subcorpus ditribution	*Steps*	English Subcorpus distribution	Spanish Subcorpus distribution
[INTRbak]	(HP: 62%)	(MP: 52%)	Est	(MP 54%)	(MP 44%)
			Pro	(-)	(-)
[INTRgap]	(HP: 76%)	(MP: 42%)	Prv AND/OR	(LP: 34%)	(LP: 10%)
			Lim	(MP: 52%)	(LP: 32%)
[INTRnew]	(C: 86%)	(C: 92%)	Obj AND/OR	(HP: 70%)	(HP: 62%)
			Pro	(LP: 32%)	(LP: 34%)
METH Moves			*Steps*		
[METHdat]	(C: 96%)	(C: 96%)	Sou AND/OR	(MP: 46%)	(HP: 72%)
			Siz AND/OR	(C: 84%)	(C: 90%)
			Crit	(HP: 72%)	(HP: 50%)
[METHproc]	(C: 96%)	(C: 80%)	App OR	(-)	(-)
			Xmp	(C: 92%)	(HP: 62%)
[METHana]	(LP: 34%)	(LP: 60%)	Cls AND/OR	(LP: 36%)	(LP: 28%)
			Proc	(MP: 48%)	(LP: 36%)
RESU Moves			*Steps*		
[RESUobs]	(C: 100%)	(C: 100%)	Gen AND/OR	(MP: 50%)	(MP: 40%)
			Spe AND	(HP: 78%)	(HP: 90%)
			Acc	(C: 90%)	(C: 66%)
[RESUnob]	(-)	(-)	Neg	(-)	(-)
CONC Moves			*Steps*		
[CONCout]	(C: 90%)	(HP: 78%)	Sig AND/OR	(C: 80%)	(HP: 70%)
			Lim AND/OR	(-)	(-)
			Int	(LP: 28%)	(LP: 16%)
[CONCres]	(HP: 66%)	(HP: 48%)	Imp OR	(MP: 54%)	(MP: 42%)
			Fth	(-)	(-)

Once the prototypical rhetorical structures for each language have been obtained, our top-down methodology moves to the next step. In this sense, most linguistic studies on research papers and abstracts have focused on their rhetorical structures; however, there are also a number of studies examining different linguistic and stylistic features (Flowerdew, 2012). In this sense, Flowerdew (2012: 1032) identifies five key grammar areas that are to be studied in order to identify prototypical grammatical resources of genres: tense usage, nominalization, voice, personal pronouns and stance markers.

In the present study, we identify some of the key areas suggested by Flowerdew so as to be able to offer syntactic and grammatical prototypical resources such as tense usage, voice and nominalization. In order to identify this lexico-grammatical resources typical of abstracts, we have used another tool that we call *Browser* and that was designed by some of the computer engineers working in ACTRES; the Browser works as a corpus manager software including the tagged rhetorical information of the abstracts "cf. Figure 2" "cf. Figure 3"; this software allows us to obtain concordances, statistics for the occurrence of moves and steps in each subcorpus, wordlists and move and/or step based phraseological information.

Figure 2: ACTRES Corpus Browser

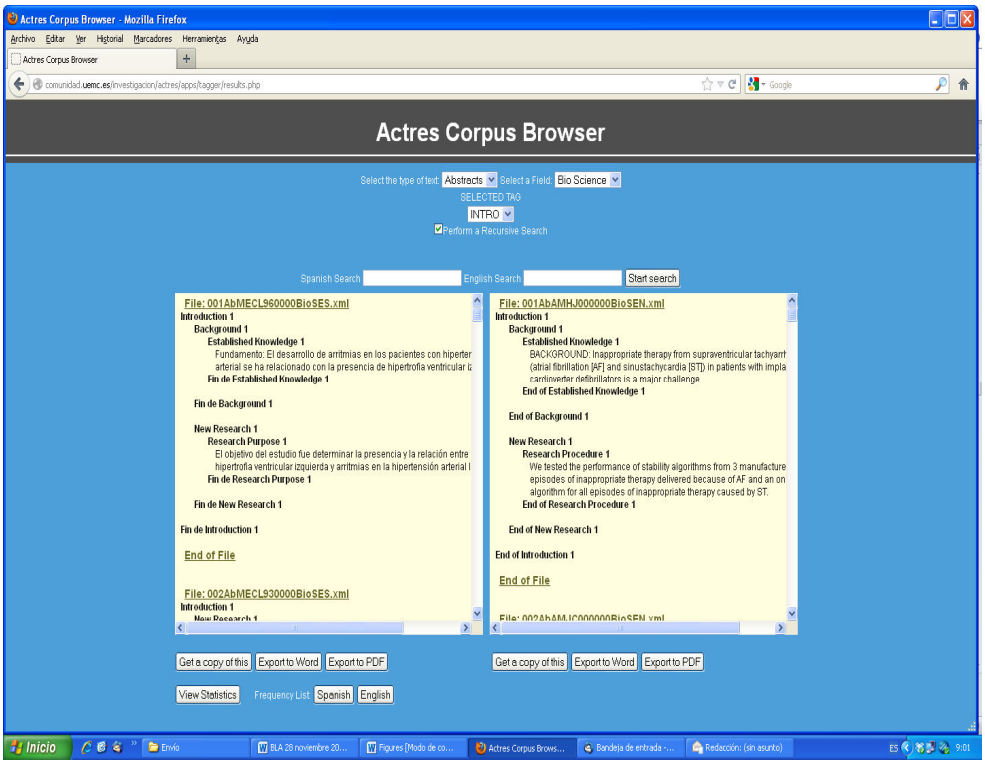
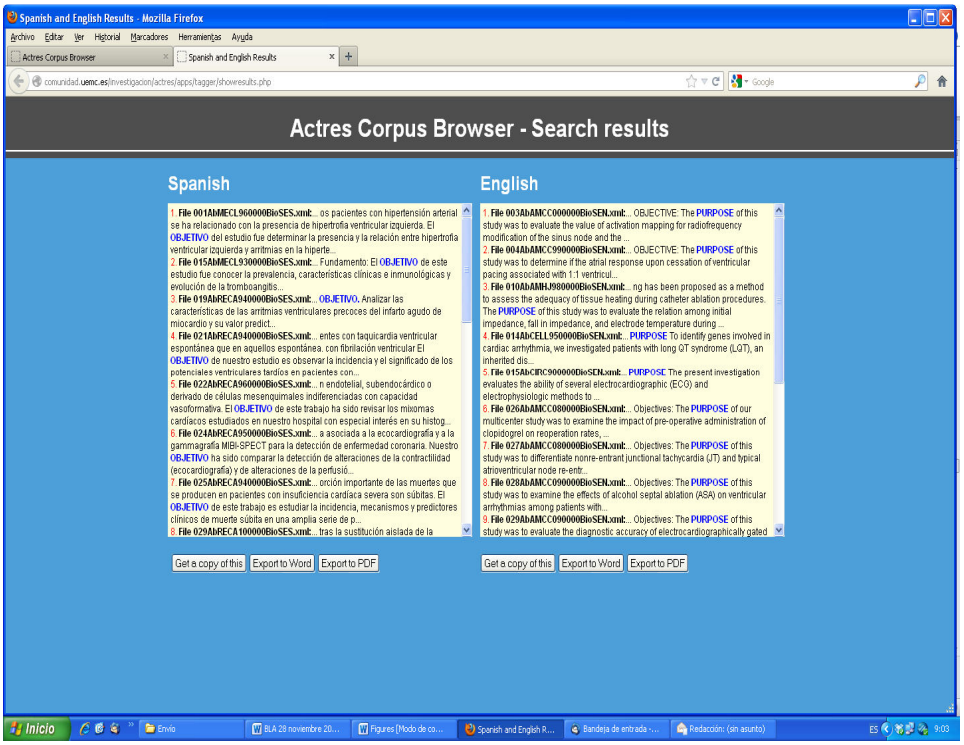


Figure 3: ACTRES Corpus Browswer (concordances)



The Browser allowed us to identify the most typical lexico-grammatical resources for each move in each section and in the two languages under study: verbs, tenses and structures occurring frequently in a particular move of the genre. For economy and convenience, only those resources identified for the most typical move in each section will be presented in this study. In the English subcorpus the aim of the research, as reported in the abstract, is expressed by the following structures "cf. Tables 2 and 3", ordered by typicality. The purpose of describing the Spanish structures is to recognize differences in the form of expressing ideas so as to make the user aware of them.

Table 2: Prototypical Resources in [INTRnew] in the English Subcorpus

	The purpose (27%) The aim (13%) The objective (2%)	of this (31%) of our (4%) of the present (4%) of the current (2%) of the (2%)	study (41%)	was to… (40%) is to… (2%)	determine …
	We (34%)			assessed, tested… (34%)	
	To describe, evaluate … (9%)				
	This study (9%)			sought to evaluate (8%)	
	This test (4%)			was developed (2%)	

The most used resources in this area are verbs. In addition to the verb *to be,* we find another 19, which can be classified into 3 main lexical groups: research verbs (to investigate, to determine, to examine, to describe) (over 70%); evaluation verbs (to evaluate, to assess); and comparison verbs (to compare, to correlate). The most typical of all are evaluate, determine, examine, investigate, summarize and assess. The verb tense is exclusively the past tense (was, analysed, assessed…). Complements are usually nouns indicating characteristics or effects (effects, incidence, prevalence…).

Spanish uses the following five typical patterns to convey 'aim' in step [INTRnewObj], listed by typicality.

Table 3: Prototypical Resources in [INTRnew] in the Spanish Subcorpus

	El objetivo (37%) El propósito (2%)	de este (27%) de nuestro (10%) del presente (4%)	estudio (27%) trabajo (10%)	es (22%) fue (12%) ha sido (4%)	analizar…
				Estudiar, analizar… (31%)	
				Se evalúa, se comenta (12%)	
				Analizamos, hemos diseñado… (10%)	
	Este estudio (4%) Este trabajo (2%)			L	

Spanish writers use a variety of verb tenses: the simple present (37%), the past (14%) and the present perfect (8%). There are 19 typical verbs; apart from *ser*, they can be grouped into three different subsets: verbs indicating research (*analizar, diseñar, determinar, examinar, conocer*…), those indicating evaluation (*evaluar, valorar*) and those indicating comparison (*comparar*).

Complements are nouns indicating concepts (83%) such as illness, factors, symptoms, characteristics, associations, etc., followed by one or several prepositional groups of varying complexity, e.g. *de / entre* (illness, factors, procedure) + *en* (patients, illness) + *con* (illness, symptoms), e.g. *la incidencia de los potenciales ventriculares tardíos en pacientes con taquiarritmias ventriculares malignas…, el grado de asociación ente adiposidad general y abdominal*….

The most typical move and step in section METH is [METHdatSou] for both subcorpora. In English, the source of the data analyzing the methods in RPs is expressed by three different structures, listed by typicality in Table 4:

Table 4: Prototypical Resources of [METHdat] in the English Subcorpus

	We analyzed We studied We identified (28%)	the + … of	symptoms	in + size of the study
	Size of the study A total of + size of the study (24%)	with + symptoms	from + source	were studied, tested, analyzed…
	The study population The group population (17%)	was/consisted/ comprised	size	who referred (42%)/underwent (42%) /had undergone

Verbs are the most important resources used in this step and can be classified into two main lexical groups: research verbs (to study, to analyze, identify) (55%) and ontological verbs (to be, to comprise, to include). The only verb tense used is the past, either in the active or the passive voice. Nouns indicating size and/or type of study (patients, group population…) (86%) or factors (illness, symptoms…) are the principal types of complements. Other significant complements are adjectives indicating the type of study carried out (controlled (14%), randomized (8%)).

Spanish uses three typical resources to express the step [METHdatSou] "cf. Table 5":

Table 5: Prototypical Resources in [METHdat] in the Spanish Subcorpus

	Se estudiaron Se analizaron (46%)	
	X pacientes El grupo de indivíduos (12%)	presentaron/ (12%) formaron (4%) lo formaron (3%)
	Analizamos (6%)	

Spanish writers combine the present and past tenses in this step, although past tenses are preferred (*se estudiaron, se analizaron* (73%). Typical verbs other than *ser* can be classified as research (*estudiar, analizar*) (>75%), evaluating (*evaluar*) and comparison verbs (*comparar*), as well as ontological verbs (*formar, incluir).*

Complements are mainly nouns indicating recipients of treatments (*pacientes, grupo de control, grupo de indivíduos*).

[RESUobsAcc] is the most typical step within the section [RESU] in the English subcorpus and exhibits three representative lexico-grammatical structures. "cf. Table 6":

Table 6: Prototypical Resources in [RESUobsAcc] in the English Subcorpus

	There were/was (30%)	(no) (significant) differences (40%)	In + results/subjects Between/among + subjects
	(No) Significant/Major (…) (24%)		Was/were observed (54%)
	Of the + number of + patients, number of patients (18%)		Received Showed

Apart from verb *to be*, verbs expressing results and or distribution are the main resource used in this step: *to show*, *observe*, *receive* (85%). The verb tense is exclusively the past. Complements are usually nouns indicating characteristics or effects (*difference(s) risk, mortality, symptoms,* etc.), followed by one or several prepositional groups of differing complexity.

Spanish utilizes three significant lexico-grammatical resources in this step. They are presented below in order of typicality. "cf. Table 7"

Table 7: Prototypical Resources in [RESUobsAcc] in the Spanish Subcorpus

	(No) se observaron /apreciaron/ identificaron (18%)	diferencias significativas	+ comparative clause	
	Se obtuvo/detectó/analizó/ identifcaron /constató /asoció (14%)		Results	
	El estudio/la enfermedad/los pacientes (8%)		Demostró (2%)/mostró (6%)	Comparative clause

Verbs expressing research in the past tense (*observaron*, *identificaron*, *detectaron*, etc.) are the most common resource in this area (>80%), as well as nouns indicating recipients (*pacientes/enfermedad*) followed by complex prepositional groups.

Finally, lexico-grammatical resources for [CONCoutSig] are illustrated in the English subcorpus with the structures shown in Table 8:

Table 8: Prototypical Resources in [CONCoutSig] in the English Subcorpus

	Type of study (28%)	Suggest(s) (12%) Show(s) (8%) Indicate (s) (4%) Reflect (s) (4%)
	(These) Finding(s) (14%) Our Finding(s) (2%)	Suggest (10%) Indicate
	Subjects (8%)	Showed (6%)

Apart from verbs indicating reference (*suggest, indicate, reflect*), the most relevant resources in this area are nouns expressing results (*findings, data, results*) (18%). Active voice and simple present or past tenses are the typical verb forms.

Table 9: Prototypical Resources in [CONCoutSig] in the Spanish Subcorpus

	Findings (resultados, servicios, prueba) (16%)	es/son (16%)	un método eficaz para (6%) frecuente (4%)
	Findings (resultados, servicios, prueba) (12%)	se desarrollan (4%), presentan, satisfice(n)	
	Hemos encontrado Encontramos (4%)	que los resultados son (4%)	

Verbs are in the simple present or present perfect. Apart from *ser* (to be) (16%) the most frequent verbs express progress (e.g. *desarrollar(se)*). The most common elements used as complements are adjectives (*útil, eficaz frecuente*) (58%).

5.2 Juxtaposition and Contrast

Juxtaposing the results of the description enables us to map the differences in prototypical structures of abstracts in the two languages, and it is essential for our study since it provides the list of the different resources used to convey a given meaning. It also gives information about whether a move or a step occurs in the two languages or not, as well as whether it occurs more frequently in one language than in the other. The data retrieved in this way from our corpus allow us to bring the prescriptive component into the process (Chesterman 1999), to ensure that "non-validated" options are excluded from the prototype (Rabadán 2008: 250).

Both sets of abstracts use the same sections, moves and steps; however their distribution is quite different for the two languages. In the Introduction section, our analysis shows that the move [INTRgap] is not very popular in Spanish (42%) compared with English (76%). In other words, Spanish authors prefer highlighting the principal findings of their own study ([INTRnew] 92%) rather than indicating the findings of other studies.

For the Material and Methods section, English and Spanish authors include a description of the materials and methods used in their research ([METHdat] 96%) as well as of the experimental procedures ([METHproc], although in the latter case, the frequency of occurrence varies between the subcorpora (96% for the English and 80% for the Spanish).

The first move of the Results section seems to be particularly important in terms of rhetorical structure, since it has 100% presence in both subcorpora.

However, the distribution of the steps shows a certain degree of variation; in this sense, English scientists focus on *accounting for the observations made* [RESUobsAcc] (90% occurrence in the English corpus against 66% in the Spanish one), whereas Spanish authors center their attention on *the description of the specific results* [RESUobsSpe] (90% against 78% in the English corpus). These data suggest that Spanish scientists are more interested in describing the results of the study, whereas the Anglophones are more concerned with their interpretation.

In the Conclusion section, English-speaking scientists concentrate their attention on the *specific research outcomes* [CONCout] (90%) and more particularly on their significance [CONCcoutSig] (80%) and implications ([CONCresImp] 54%). These aspects are apparently not as important for Spanish authors (70% significance and 42% implications).

Turning to the lexico-grammatical resources, both English and Spanish authors use the same five structures to express the aim of the paper [INTRnew], although the two languages exhibit different typicalities. The most prototypical in both is the syntagm including the words *el objetivo de este estudio – the purpose / aim of this study* followed by *is to...* However, in the Spanish Introductions the simple present is more common than the past, which is almost the only possibility in English.

In the same move, the non-finite and finite purpose clauses with first person plural subjects are more frequent in Spanish (31% vs. 11%), while English prefers finite purpose clauses (42% vs. 39%). Personification is used in both languages (for example, *this study evaluated*, in the English subcorpus), but with similarly low frequencies: 9% in the English subcorpus and 6% in Spanish. Finally, the lexical units which serve as complements for the verbs in English as well as in Spanish, indicate effects, although Spanish also uses nouns referring to recipients.

Lexico-grammatical resources for [METHdat] differ between languages. Whereas English authors prefer active voice structures (28%), Spanish writers show a preference for the passive in this move (39%).

On the other hand, although both languages include a step to express the size of the study, the lexico-grammatical resources used are different; English uses a passive voice structure (24%) and Spanish uses the active (12%).

Other differences can be found in verb tenses: English only uses the past while Spanish combines simple present (26%) and past tenses.

Apart from nouns indicating recipients, regular patterns for complements in Spanish have not been found in the present study. English uses nouns for recipients but also for factors, as well as adjectives qualifying the type of study carried out (22%), such as *randomized, controlled,* etc.

Both languages use personification in [RESUobsAcc] although its distribution is different, as English uses it more frequently than Spanish (18% vs. 8%). No other significant differences can be found in [RESUobsAcc] between the two languages. It should be mentioned that this section shows a preference for variation in terms of expression rather than for recurrent patterns as was the case in INT and METH.

Finally, both English and Spanish authors use one out of three similar structures to express the conclusion of the paper [CONCoutSig].

5.3 Results

The results of the contrastive analysis allow for the development of a prototype of the rhetorical and lexico-grammatical elements of scientific abstracts in each of the languages. For usability reasons, our basic prototype only includes three lexico-grammatical options for each move at the moment, although versions with more complex structures will be available in the future. Although in the present paper we have only described the lexico-grammatical resources for one of the moves in each section, in the table below all three options for each move are shown. These options have been obtained using the methodology described in the paper and are organized in order of typicality "cf. Table 10".

This information in the table can be considered as constituting a sort of template featuring the correct and acceptable options that are most frequently used by the global scientific discourse community, and will be fed into the *Scientific Abstract Generator* interface.

"see Section 6" This template brings the necessary prescriptive element into play and is useful insofar as it gives answers to the problems of end-users.

Table 10: Template for Abstracts Rhetorical Structure

Title	Noun (Object of study) + in + type of subjects + with + features
	A + study + for + features
	Evaluation of + noun (object of study) + on + noun (group) (pathology)
[INTRbakEst]	Noun + is well established (other organizing verbs)
	Noun + has been receiving increased attention
	Noun + is a major challenge
[INTRgapLim]	Noun + has/have not been well (other adjectives) + described (other description verbs)
	However, studies have been limited to (other description verbs)
	Noun (subjects) have not been adequately (other adverbs) evaluated (other evaluation verbs).
[INTRnew Obj]	The aim of this study was to investigate (other research verbs) + noun (group) (object of study)
	The purpose of this study was to determine (other identification verbs) + noun (group) object of study
	We assessed (other evaluation verbs) + noun (object)
[METHdat Sou] **[METHdat Siz]** **[METHdat Crit]**	We analyzed (other research or ontological verbs) + the + noun + of + symptoms + in + size of the study (x patients)
	A total of + size of the study with + noun (group) (symptoms) + were studied
	The study population consisted of + size of the study + who referred + noun (group) (criteria for inclusion)
[METHprocXmp]	A + noun (group) (type of study) + was performed (other evaluative or distribution verbs) by using + procedure + in + number subjects
	Feature + was assessed (other evaluative verb) by + procedure
	Noun (group) (Procedure or apparatus) was used to + infinitive
[RESUobsGen]	Noun (group) (Procedure or apparatus) + demonstrated + noun + in number (of) subjects
	(Successful) (other adjective) elimination of + noun (group) (pathology) + was achieved + in + number (of) subjects
	Noun (group) (Results) did (not) affect + noun group
[RESUobsSpe]	There were (no) significant differences in + Noun (group) (results) + among + users
	There was (no) major) + noun (group) (results) + in + subjects

	(No) significant differences in + noun (group) (results) + were observed
[RESUobsAcc]	
[CONCout Sig]	Noun (group) (Type of study) + demonstrated (other evaluative or distribution verb) + (that) + noun (clause)
	Noun (group) (Results) + indicate + Noun (group) is increased by + noun (group)
	Noun (group) (Subject of study) + showed (other evaluative or distribution verb) increase/decrease + in + noun (group)
[CONCresImp]	In conclusion + noun (group) (object of study) is/is not + (only) + effective (other adjective) for noun (group) + with + noun (group) (features)
	We + conclude + that + noun (group) + is +
	Noun (group) (Object of study) + proved of + (un/limited) (other adjective) value + to

6. The Application Prototype: the Scientific_Abstract_Generator

The *Scientific_Abstract_Generator* (hereafter called *the Generator*) is an online interface for Spanish-language scientists in need of language aid when writing their abstracts in English.

It has been built using Hypertext Markup Language (html), Javascript and Hypertext Preprocessor (PHP). It consists of textual plus phraseological features, which are to be used simultaneously at every writing stage. For usability reasons, the interface is displayed on a single 'screen' featuring sections corresponding to the moves and/or steps identified in the textual template. In the drop-down menus the three phraseological options mentioned above are displayed. This overall design makes the Generator user-friendly. The content solves problems of how to string words together correctly and acceptably, both aspects identified by prospective users as problematic. "see Section 1." As shown in Fig. 4 "cf. Figure 4", the interface guides the user in Spanish through the different sections, taking into account the fact that our targeted user 'starts from Spanish and tries to fill in the slots with formally equivalent resources', as reported in section 1. "see Section 1".

The instructions are simple and straightforward, and include an indication of how many words are acceptable for the length of each move. A prototypical example in English acts as an instant demo of what and how should be inserted in each move.

This is followed by the drop-down menu of *Sugerencias* (suggestions), which includes three phraseological options from which to choose. Once this choice is made according to the user's requirements, he has to complete the chosen string. To help at this stage, the *Generator* includes further word and phrase information that pops up on the screen as the user types the first two letters of his or her choice in Spanish when trying to complete the English string. "cf. Figure 5". An English option will then appear, and, simply by clicking on it, the selected English option will be added to the text in progress. Since the Generator has been designed for Spanish-language users the tool is unidirectional into English and works on the basis of translation equivalents. Unlike more general technical glossaries, it includes semi-technical and general language items that have acquired special uses in bioscience texts, together with all the corpus-based examples. This writing aid has been conceived as a production-oriented tool, but it may also have remedial and pedagogical uses (Hannay, 2003: 145).

Once all the sections have been completed in this fashion, it takes just a simple click and the *Generator* will automatically deliver a correct, complete abstract written on the basis of the user's choices, which can be saved in two different formats (pdf or doc).

Figure 4: Scientific Abstract Generator Interface

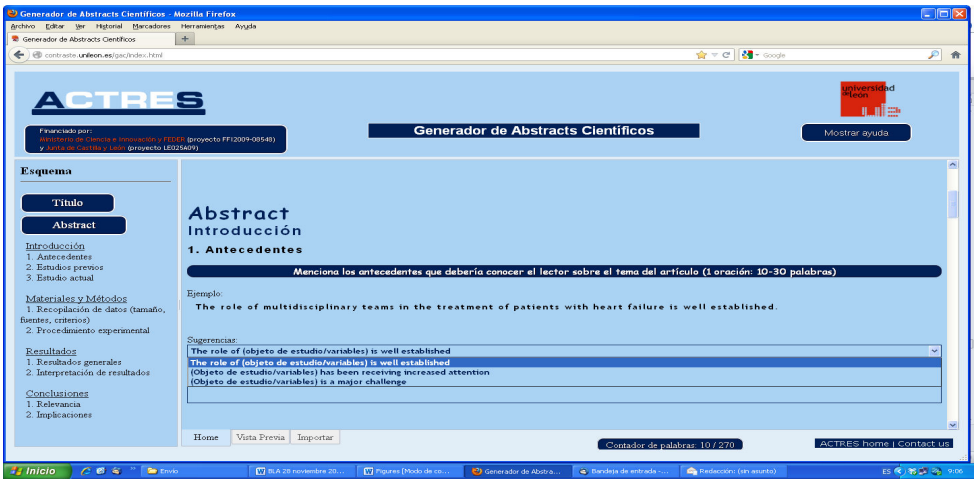
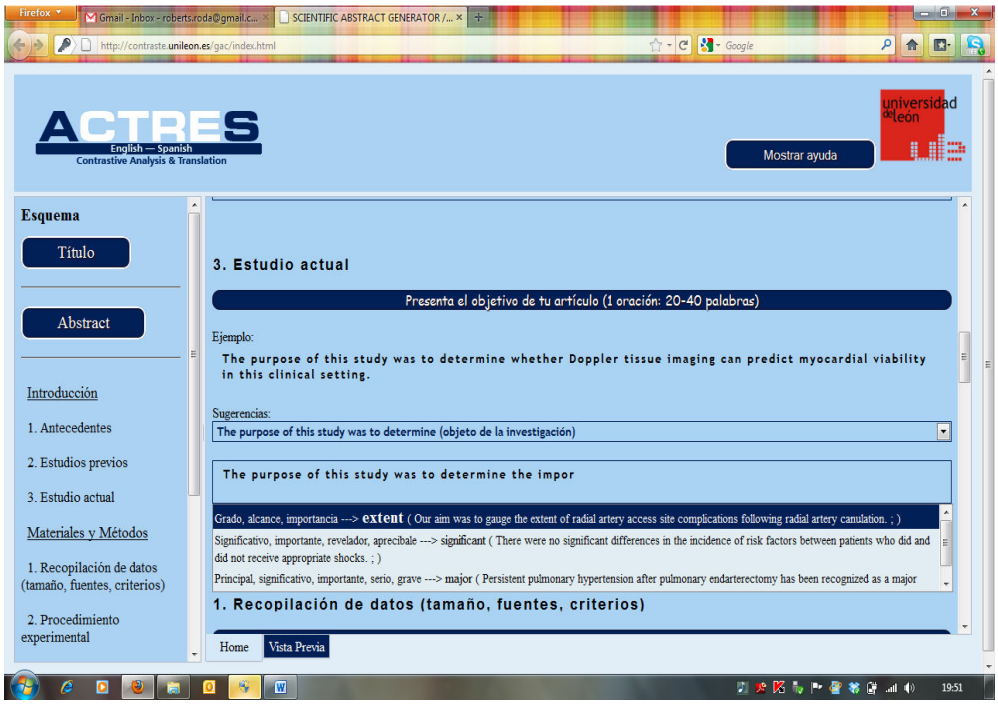


Figure 5. Lexical component of the Scientific Abstract Generator



7. Conclusion

The prototype we have presented here is the result of collaborative work between linguists and computer analysts from the ACTRES research group, which aims at producing innovative solutions for real communication problems.

The problem addressed here is a very real one and affects a considerable number of non-English-speaking professional groups and individuals when they try to report their work to their particular international discourse community. The solution envisaged is a computer-based text generator that uses reliable linguistic information suited to the needs of the end-users. By making text production easier and more error-free, this solution would hopefully contribute to professional advancement and cost reduction.

In this particular case, the solution offered is the *Scientific_Abstract_Generator*, computer software that will enable a Spanish-speaking scientist lacking advanced writing skills in English to produce a textually acceptable and linguistically correct English abstract.

Innovative features of this proposal are the use of empirical data obtained from the BioABSTRACTS_C-ACTRES corpus to identify prototypical features, the research protocol that integrates well-grounded contrastive work and evaluative procedures in order to formulate a useful, correct text template, and a clean, user-friendly computer interface design. All of these features can be replicated and transferred to other 'bad writing' situations.

Although the *Generator* is still at an early stage of development, and refinement and large-scale testing is still underway, the results of pilot testing may have an impact on the way the work of non-English-speaking scientists and academics is presented globally by making cross-linguistic written communication more efficient and more affordable.

8. Appendix 1: Key to the Rhetorical Labels Used in the Present Study

Introduction: [INTR]

- BACKGROUND KNOWLEDGE [INTRbak]
    - Established knowledge in the field [INTRbakEst]
    - Main research problems [INTRbakPro])
- INDICATING A GAP [INTRgap]
    - Previous studies [INTRgapPrv]
    - Limitation of previous research/studies [INTRgapLim]
- NEW RESEARCH [INTRnew]
    - Research purpose [INTRnewObj]
    - Main Research procedure [INTRnewPro]

Materials and methods: [METH]

- DATA COLLECTION PROCEDURE [METHdat]
    - Source of data [METHdatSou])
    - Data size [METHdatSiz]
    - Criteria for collection [METHdatCrit]
- EXPERIMENTAL PROCEDURE [METHproc]
    - Experimental process [METHprocXpm]
    - Research apparatus [METHprocApp]
- DATA-ANALYSIS PROCEDURE [METHana]

- Data classification [METHanaCls]
- Instrument procedure [METHanaProc]

Results: [RESU]

- CONSISTENT OBSERVATION [RESUobs]
  - Overall observation [RESUobsGen]
  - Specific observation [RESUobsSpe]
  - Accounting of observation made [RESUobsAcc]
- NON-CONSISTENT OBSERVATION [RESUnobs]
  - Negative results [RESUnobsNeg]

Conclusion: [CONC]

- SPECIFIC RESEARCH OUTCOME [CONcout]
  - Indicate significance [CONcoutSig]
  - Interpret research results [CONcoutInt]
  - Limitation of present research [CONCoutLim]
- RESEARCH CONCLUSION [CONCres]
  - Implications [CONCresImp]
  - Further studies [CONCresFth]

9. References and Webgraphy

Aston, G. (1999). Corpus use and learning to translate. Textus, 12, 289-314.

Austermühl, F. (2001). Electronic tools for translators. Manchester: St. Jerome.

Bhatia, V. K. (1993). Analysing genre: Language use in professional settings. London: Longman.

Bhatia, V. K. (2004). Worlds of written discourse. A genre-based view. London: Continuum.

Biber, D. (1993). Representativeness in corpus design. Literary and linguistic computing, 8, 243-57.

Biber, D., Conrad, S. & Reppen. R. (Eds). (1998). Corpus Linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, D., Connor, U. & Upton, T.A. (Eds.). (2007). Discourse on the move. Using corpus analysis to describe discourse structure. Antwerp: John Benjamins.

Bowker, L. & Pearson, J. (2002). Working with specialized language: A practical guide to using corpora. London: Routledge.

Chesterman, A. (1998). Contrastive functional analysis. Antwerp: John Benjamins.

Chesterman, Andrew. (1999), The empirical status of prescriptivism. Folia translatologica, 6, 9-19 [Online] Available: http://www.helsinki.fi/~chesterm/1999aPrescr.html (November 28, 2012).

Cognetics Corporation. (2011), Usability testing. [Online] Available: http://www.leadersintheknow.biz/UsabilityUserExperience/UsabilityTesting/tabid/176/Default.aspx (November 28, 2012).

Council of Europe. (2000), The common European framework of reference for languages. European language portfolio [Online] Available: http://www.coe.int/t/DG4/Portfolio/?M=/main_pages/levels.html (November 28, 2012).

Flowerdew, J, & Dudley-Evans, T. (2002). Genre analysis of editorial letters to international journal contributors. Applied linguistics, 23, 463-489.

Flowerdew, L. (1998). Corpus linguistic techniques applied to textlinguistics. System, 26, 541-552.

Flowerdew, L. (2005). An integrated approach of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticism. English for specific purposes, 24, 321-332.

Flowerdew, L. (2008). Corpus-based analysis of the problem-solution pattern. Antwerp: John Benjamins.

Flowerdew, L. (2012). Grammar and the research article. In C. Chapelle (Ed.), The Encyclopedia of Applied Linguistics. London: Wiley-Blackwell.

Foraker Design, (2002-2010), Usability first. Introduction to usability. [Online] Available: http://www.usabilityfirst.com/intro/index.txl. (January 2010).

Gläser, R. (1995). The LSP genre abstract revisited. In R. Gläser (Ed.), Linguistic features and genre profiles of scientific English (pp. 97-105). Frankfurt: Verlag.

Hannay, M. (2003). Types of bilingual dictionaries. In P.T. Van Sterkenburg (Ed.), A practical guide to lexicography (pp.145-153). Antwerp: John Benjamins.

Harris, B., & Hofmann, T.R. (1973). Formulated abstracting: An experiment in regularized content description. Paper presented at the First open conference of information science, Canada, 98-107.

Henry, A., & Roseberry, R.L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: Letter of application. English for specific purposes, 20, 153-167.

Kreitzberg, Charles & Little Ambrose. (2009), Usability in practice: Useful, usable and desirable. Usability as a core development competence. MSDN magazine. [Online] Available: http://msdn.microsoft.com/en-us/magazine/dd727512.aspx (January 2010).

Krzeszowski, T. P. (1990). Contrasting languages. The scope of contrastive linguistics. Berlin: Mouton de Gruyter.

Labrador de la Cruz, B. (2003). Typicality in the use of English and Spanish indefinite quantifiers: A practical approach, Babel, 49, 41-58.

Landauer, T. K. (1995). The trouble with computers: Usefulness, usability, and productivity. Cambridge, MA: The MIT Press.

López Arroyo, B., & Méndez Cendón, B. (2007). Describing phraseological devices in medical abstracts: an English/Spanish contrastive analysis. Meta, 52, 503-516.

López Arroyo, B., Fernández Antolín, M.F. & de Felipe Boto, M.R. (2007b). Contrasting the rhetoric of abstracts in medical discourse. Implications and applications for English Spanish translation. Languages in contrast, 7, 1-28.

Lorés Sanz, R. (2003). On the rhetorical structures of abstracts. In G. Luque & A. Bueno (Eds.), Languages in a global world (pp. 73-80). Jaén: Universidad de Jaén.

Medscape [Online] Available http://www.medscape.com/ (April 11, 2014).

Méndez-Cendón, B., & López Arroyo, B. (2003). Intralinguistic analysis of medical research papers and abstracts. Rhetorical and phraseological devices in scientific information. Terminology, 9, 247-269.

Nord, Ch. (1997). Translating as a purposeful activity. Manchester: St. Jerome.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. English for specific purposes, 16, 119-138.

Quesenbery, Whitney. (2001), What Does Usability Mean: Looking Beyond 'Ease of Use'. [Online] Available http://www.digitalspaceart.com/projects/cogweb2002v2/papers/whitney/whitney1.html (May 6th 2014)

Rabadán Álvarez, R. (2002). Análisis contrastivo y traducción especializada inglés-español: el programa ACTRES. In J. Mª. Bravo Gozalo (Ed.), Nuevas perspectivas de los Estudios de Traducción (pp. 35-56). Valladolid: Universidad de Valladolid.

Rabadán Álvarez, R. (2007). Divisions, description and applications- The interface between DTS, corpus-based research and contrastive analysis. In Y. M. Gambier, Schlesinger, M. & Stolze, R. (Eds.), Doubts and directions in Translation Studies (pp. 237-252). Antwerp: John Benjamins.

Rabadán Álvarez, R. (2008). Refining the idea of 'applied extensions. In A. Pym, Schlesinger, M. & Simeoni, D. (Eds.), Beyond descriptive Translation Studies: Investigations in homage to Gideon Toury (pp. 103-118). Antwerp: John Benjamins.

Rabadán Álvarez, R. (2008a). Tools for English-Spanish cross linguistic applied research. Journal of English Studies, 5-6, 309-324.

Rabadán Álvarez, R. (2010). Applied Translation Studies. In Y. Gambier & Doorslaer, L.V (Eds.), Handbook of Translation Studies. Volume 1, [HTS 1] (pp. 7-11). Antwerp: John Benjamins.

Russell, P. 1988 . How to write a précis. Ottawa: University of Ottawa Press.

Sager, J.C., Dungworth, D. & Mcdonald, P. (1980). English special languages. Bransletter.

Scott, M. (1996). WordSmith tools version 4.0. Oxford: Oxford University Press.

Suter, Hans J. (1993). The Wedding Report. A prototypical approach to the study of traditional text types. Antwerp: John Benjamins.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

Tribble, Ch. (2001). Small corpora and teaching writing. In M. Ghadessy, Henry, A. & Rosberry, R. (Eds.), Small corpus studies and ELT (pp. 381-405). Antwerp: John Benjamins:.

Upton, T. & Connor. U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. English for specific purposes, 20, 313-329.

Upton, T. (2002). Understanding direct mail letters as a genre. International journal of corpus linguistics, 7, 65-85.

Vergaro, C. (2004). Discourse strategies of Italian and English sales promotion letters. English for specific purposes, 23, 181-207.