WRITING ABSTRACTS: TECHNOLOGICAL APPLICATIONS FROM A CORPUS-BASED STUDY[1]

Rosa Rabadán
Ángeles Díez
Ramón-Ángel Fernández
Universidad de León (Spain)

Belén López
Universidad de Valladolid (Spain)

ABSTRACT: Abstracts, which constitute a secondary genre based on the Research Paper (RP), have often been analyzed in order to observe how information has been rendered for translation or contrastive analysis purposes. However, in this genre, as in many others, "while there is a wealth of descriptive research, generally speaking, the information is not directly amenable to applied endevours" (Rabadán, 2008: 103). The aim of this paper was to describe the methodology and the tools devised by the ACTRES research group to bridge the transition between linguistic description and procedural information. The first step of this process was to design a small special corpus of scientific abstracts, the BioAbstracts_C-ACTRES. The macro and microlinguistic characteristics of this corpus were analyzed in order to find the most prototypical rhetorical, grammatical and lexical features of this genre. Then, we identified the "anchors" (Rabadán: in press) relevant for the native speakers of Spanish. Finally, a prototype of a writing application, the *Scientific_Abstract_Generator,* has been designed. Still under development, it aims at helping native Spanish users who are non-linguist field experts, to write scientific abstracts in English.

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]. The acronym stands for *Análisis contrastivo y traducción English-Spanish: Aplicaciones II*/"Contrastive analysis and translation English-Spanish: Applications II".

INTRODUCTION

For a scientist presenting his/her work sucessfully to the research community is a top priority. If it implies overcoming difficulties in cross-linguistic written communication it can also become a distressing experience. ACTRES (http://actres.unileon.es) research fills in a niche which has been neglected in other research programs and tries to cater for a pressing need of non-linguist users: to make available English-Spanish bilingual aids for written communication addressed to speakers of Spanish as a first language.

These writing aids are envisaged as user-friendly computer applications termed 'generators' that will enable non-linguist users to make correct decisions on the basis of validated corpus-based contrastive research. These applications will consist of a useful and usable interface giving access to i) textual-linguistic guidelines and, ii) terminological information.

This paper sets out to present the tools and the methodology used to obtain the empirical data that will feed our application prototype. We will concentrate on abstracts, a type of metatext that functions as the first and foremost introductory tool of research work in science.

METHODOLOGY

In the design of this application prototype the linguists' task has consisted in completing the first of the phases in the construction process of a text in natural language for the communication of specific puposes: the planning (Jordán, 1992: 7). Subsequently, the designer has been provided with the linguistic information necessary for the second phase: the generation itself.

During the planning we have analyzed the meaning, use and function of abstracts (López et al., 2007: 10), because, as Baker (1993: 237) states, correspondence in meaning amounts to correspondence in use. Therefore, our methodology is applied using the most useful tool we have in linguistics: a computerized corpus, which allows us to describe real utterances within a communicative situation.

We built a specific purpose corpus, according to pragmatic criteria. In selecting our texts, we considered the representativity and availability of the abstracts; in other words and according to Nwogu (1997: 121), the abstracts were

chosen to ensure a representative sample of the language of members of the discourse community. Availability, on the other hand, refers to the ease with which abstracts constituting the corpus can be obtained.

As we are interested in the acceptability of abstracts by the other language discourse community, we built a corpus, the BioAbstracts_C-ACTRES, which can be described as comparable, bilingual, synchronic and annotated; first, because it is based on fifty abstracts originally written in English and fifty in Spanish; second, because the sample texts were chosen by publication date; particularly, only those abstracts published in the last decade were considered for inclusion in the corpus; and third we marked the rhetorical structure of every abstract in order to describe their similarities and differences in their construction.

In order to compile the English comparable corpus we started our search on the Internet; and in this sense, several Internet sites contain links to scientific community databases. However, in a second stage, we restricted our search to those robust search engines such as Medscape selection, for instance, which select abstracts and Research Papers depending on their scientific validity, importance, originality and contribution to the scientific community, in the case of the example we are offering, to the medical specialty. Medscape selection criteria could be called into question; nevertheless, each title included in it has to meet one of the following criteria:

F) expert opinion of pre-eminent clinicians and researchers (…);

G) named as one of the nine English-language international general medical journals whose full-time editors are members of the International Committee of Medical Journal Editors;

H) inclusion on a 1994 internal JAMA (Journal of American Medical Association) journal list;

I) a journal impact factor greater than 2 as ranked by the Institute for Scientific Information's Journal Citation Reports;

J) and high readership scores determined by PERQ (Pharmaceutical and Health Care –related promotion research) and published on Medscape.

Further criteria we used affected the journal impact. In this sense, the Institute for Scientific Information (ISI) ranks journals according to their impact in the scientific community and this is the main criteria used for our corpus compilation. Only those abstracts published in journals with greater impact were selected.

Regarding the Spanish comparable corpus, international impact could not be used; however, those journals included in the ISI were chosen.

As for the procedure for analyzing the corpus Bhatia (2004) proposes a comprehensive procedure: the multidimensional and multi-perspective research methodology. Its basis is the study of texts from three complementary viewpoints: the textual, the socio-cognitive and the social space. According to this author this kind of analysis should account for a combination of *text-internal* and *text-external* features, such as rhetorical, cognitive and lexico-grammatical elements, and text production and interpretation by their discursive communities respectively.

The approach adopted in this study is purely generic in its first stage. The texts are analysed basically within their socio-cognitive space, that is to say, as communicative events expressed through rhetorical resources. In a second phase, with applied purposes, we have also dealt with the textualization of certain lexico-grammatical features, i.e. the analysis of the textual space.

The Socio-Cognitive Space of the Bio_Abstracts_C-ACTRES Corpus

Abstracts are defined by ISO 214-1976 (E), as an "abbreviated, accurate representation of the contents of a document, without added interpretation or criticism and without distinction as to who wrote the abstract" (Gläser, 1995: 97); that is to say, this type of abstract has been derived from a fully elaborated text by condensing its relevant information, the RP.

There are two basic types of abstracts**,** informative or RP abstracts, and descriptive**.** RP abstracts constitute a well-defined genre with definite attributes and a unique style; it has to be brief, accurate, objective, complete, and intelligible, and it has to be presented in the same format of the RP in order to facilitate the skimming of the RP. 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).

Sometimes, informative abstracts are the only piece of writing that readers can read. Thus, they have become a key to the content of the whole text. Moreover, because several journals publish only abstracts as a source of quick information and orientation, in some cases, the informative abstract is the only piece of published writing. Therefore, a well-written abstract becomes increasingly important in directing readers to articles of potential value (López et al., 2007: 8).

Our corpus shows that the informative scientific abstracts analyzed include the four sections, which are divided into moves ("meaningful units realized by linguistic means which fulfil a communicative function") (Biber *et al.*, 2007: 23) and steps (small rhetorical units moves can be divided into) (see Table 1). The different combinations of sections, moves and steps compose the *rhetorical structures*.

Table 1. Rhetorical elements of scientific abstracts and their most prototypical rhetorical structure (in bold).

Section 2: **INTRODUCTION**	
Moves	*Steps*
Background information (HP)	**Established knowledge in the field** OR **(MP)**
	Main research problems (-)
Review related research (HP)	Previous research AND/OR (LP)
	Limitations of previous research (MP)
New Research (C)	**Research Purpose** AND/OR **(HP)**
	Main research procedure (LP)
Section 2: **MATERIALS AND METHODS**	
Moves	*Steps*
Data collection procedure (C)	**Source of data** AND/OR **(MP)**
	Data size AND/OR **(C)**
	Criteria for data collection (HP)
Experimental procedure (C)	Research apparatus OR (-)
	Experimental process (C)
Data-analysis procedure (LP)	Data classification AND/OR (LP)
	Analytical instrument/procedure (MP)
Section 3: **RESULTS**	
Moves	*Steps*
Consistent observation (C)	**Overall observation** AND/OR **(MP)**
	Specific observation AND **(HP)**
	Accounting of observation made (C)
Non-consistent observation (-)	Negative results (-)
Section 4: **CONCLUSION**	
Moves	*Steps*
Specific research outcome (C)	**Indicate significance** AND/OR (C)
	Limitations AND/OR (-)
	Interpret (LP)
Research conclusions (HP)	**Implications** OR **(MP)**
	Further research (-)

This qualitative analysis is accompanied by a quantitative one (Upton & Connor, 2001). Following Suter's quantitative approach (1993: 119) the moves and steps of our corpus are classified as:

6) compulsory moves and steps (C): appearing in between 100% and 80% of the moves;

7) high priority moves and steps (HP): between 80% and 60% of the moves;

8) medium priority moves and steps (MP): between 60% and 40% of the moves;

9) low priority moves and steps (LP): between 40% and 20%;

10) occasional moves and steps (-): appearing in less than 20%.

C, HP and MP in English and Spanish, combined as in the most frequent rhetorical structures (see Table 1, in bold), have been selected for the design of the first version of the *Scientific_Abstract_Generator.*

The main principle underlying this choice is one of the most relevant features of genre, its *prototipicality*, i.e. the conventional character of its texts, the regularities affecting its discursive structure (at a macrotextual level) as well as its lexico-grammatical characteristics (at a microtextual level). The individuals of a discourse community, who have a prototypical image of it, are able to associate each text to a certain prototype thanks to the recurrence of the intra and extratextual elements. Hence, the choice of the most prototypical elements is the basis to build the generator.

The Textual Space of the BioABstracts_C-ACTRES Corpus

After analyzing the socio-cognitive space we have studied the lexical, grammatical and syntactic elements that compose the textual space of our corpus. For the lexical data we have used *WordSmith* version 4.0 (Scott, 1996), a software kit that contains *WordList* to elaborate frequency and alphabetical word lists, and *Concord* to place the search word in its contexts.

We isolated those linguistic structures and elements which were representative in terms of frequency from the following types: clause type, lexico-grammatical characteristics of clause elements (subjects, verb features, complement types…) and relevant semantic features (technical terms,

subtechnical words…). For example in the Introduction for the move *New Research* we were able to isolate four different structures to express this semantic unit (see Table 2).

Table 2. Lexico-grammatical analysis of Introduction: New Research.

<table>
<tr><td rowspan="3">1. The aim of</td><td>our current/ present</td><td rowspan="3">study</td><td rowspan="3">was</td><td rowspan="4">to</td><td rowspan="4">[infinitive]
- Research verbs:
investigate,
determine,
examine, identify,
establish…;
- Evaluative verbs:
evaluate, assess,
test, measure;
- Comparative verbs:
compare.</td><td rowspan="4">the</td><td rowspan="4">[noun]</td></tr>
<tr><td>the current/ present</td></tr>
<tr><td>this/ the</td></tr>
<tr><td>2. The purpose of</td><td>this/ the</td></tr>
<tr><td colspan="3">3. The / Our aim</td><td rowspan="2"></td></tr>
<tr><td colspan="3">4. We</td><td>aim</td></tr>
</table>

THE APPLICATION PROTOTYPE: *SCIENTIFIC_ABSTRACT_GENERATOR*

Scientific_Abstract_Generator is an application prototype devised for its on-line use. It has been built by using html, Javascript and Php. It consists of a textual and a lexicographic module, which can be used simultaneously at every writing stage.

The Textual Module of the *Scientific-Abstract_Generator*

The textual module consists of a combination of drop-down menus for each move plus a *Help* section.

To use the generator the writers have to click first on the section and then on the move (see Fig.1), where they are offered a drop-down menu with several lexico-grammatical options and authentic examples from the corpus. In addition it includes some writing guidelines, which appear in the *Help* next to each move**.** As the prototype is aimed at non-linguists, the guidelines need to be easy to follow for experts in fields other than Linguistics and/or Computing.

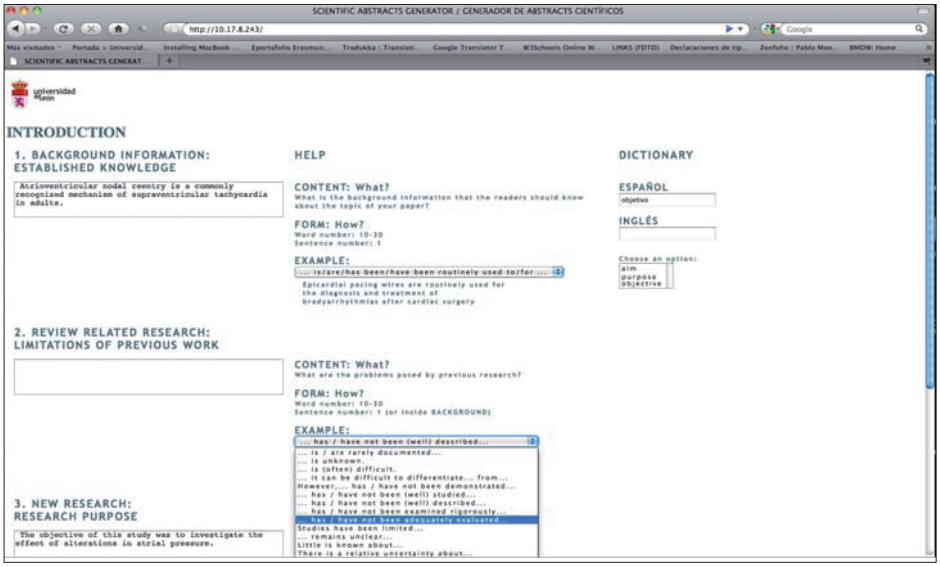

Figure 1. Components of the *Scientific_Abstract_Generator*.

The Lexicographic Module of the *Scientific_Abstracts_ Generator*

For the elaboration of the glossary (see Fig. 1), we have followed Yong and Peng's proposal (2007). These authors consider the lexicographic work as a communicative task incorporating the user in the dictionary general configuration.

Since it is designed for native Spanish users it is an electronic unidirectional bilingual Spanish-English glossary that uses translation equivalents. It is conceived of as a production-oriented tool with pedagogical purposes (Hannay, 2003: 145), to help users to write a specific textual genre. It is also specialized, because it only focuses on the terms necessary to deal with a specific matter (Bowker, 2003: 154). However, we have not separated technical from semitechnical terms. Hence, in this study independently of their origin –whether they belong to a specialized language or they are used in general language–, those lexical units carrying out a specialized and restrictive meaning were considered candidates to be included in the lexicographic module.

CONCLUSION

The prototype we have presented is the result of collaborative work between linguists and computing engineers, which constitutes the core endeavor of the ACTRES research group.

An innovative feature is the use of empirical data obtained from the BioAbstracts_C-ACTRES corpus. English and Spanish data have been analysed for prototypical features following Bhatia (2004) and contrasted in order to identify cross-linguistic 'anchors' (Rabadán: in print). The information gathered has been used to feed our 'generator' prototype with grammatical choices, terminological expertise and rhetorical guidelines that work in conjunction with basic, useful and usable computer tools. Thus, a Spanish-speaking scientist lacking expert writing skills will be able to produce a linguistically acceptable and correct abstract in English.

Forthcoming work will concentrate on the anchors so as to further improve the usefulness and applicability of our analysis.

Although the generator is still at an early stage of development, and refinement and testing is still pending, it will have an impact in both the ways Spanish science is presented globally by making cross-linguistic written communication more efficient and more affordable.

REFERENCES

BAKER, M. (1993): Corpus Linguistics and Translation Studies. Implications and applications. In M. BAKER & E. TOGNINI BONELLI (eds.). *Text and Technology* (pp. 233-250). In Honour of John Sinclair. Antwerp: John Benjamins.

BHATIA, V. K. (2004). *Worlds of Written Discourse. A genre-based view*. London: Continuum.

BIBER, D.; CONNOR, U. & UPTON, T. A. (eds.) (2007). *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Amsterdam/Philadelphia: John Benjamins.

BOWKER, L. (2003). Specialized lexicography and specialized dictionaries. In V. STERKENBURG. *A Practical Guide to Lexicography* (pp. 154-164). Amsterdam/Philadephia: John Benjamins.

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

HANNAY, M. (2003). Types of bilingual dictionaries. In V. STERKENBURG. *A Practical Guide to Lexicography* (pp. 145-153). Amsterdam/Philadelphia: John Benjamins.

JORDÁN, A. G. (1992). Lenguas y tecnologías de la información. *Centro Virtual Cervantes*. Retrieved August 15, 2008, from http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_ajordan.htm.

LÓPEZ, B.; FERNÁNDEZ, M. & DE FELIPE R. (2007). Contrasting the rhetoric of abstracts in medical discourse. implications and applications for English Spanish translation. *Languages in Contrast*, 7(1), 1-28.

LORÉS SANZ, L. (2003). On the rhetorical structures of abstracts. In G. A. LUQUE, A. BUENO & G. TEJADA. *Languages in a global world* (pp. 73-80). Jaén: Universidad de Jaén.

NOWGU, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes*, *16*(2), 119-138.

RABADÁN, R. (2008). Refining the idea of 'applied extensions'. In A. PYM, M. SCHLESINGER & D. SIMEONI. *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury* (pp. 103-118). Amsterdam/Philadelphia: John Benjamins.

RABADÁN, R. (in press). Applied Translation Studies. In Y. GAMBIER & L. VAN DOORSLAER. *Handbook of Translation Studies*. Volume 1, [HTS 1] (pp. 7-11). Amsterdam/Philadelphia: John Benjamins.

SCOTT, M. (1996). *WordSmith Tools (version 4.0)*. Oxford: Oxford University Press.

SUTER, H. J. (1993). *The Wedding Report. A prototypical approach to the study of traditional text types*. Amsterdam/ Philadelphia: 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*(4), 313-329.

YONG, H. & PENG, J. (2007). *Bilingual Lexicography from a Communicative Perspective*. Amsterdam /Philadelphia: John Benjamins.

SECTION 4

TECHNOLOGY AND LANGUAGE CORPUS