Understanding scientific communication through the extraction of the conceptual and rhetorical information codified by verbs*

Clara Inés López Rodríguez

Both translation and terminology can benefit from English for specific purposes and corpus linguistics in the study of the process of extracting information from texts, in particular, medical abstracts complying with the IMRAD (Introduction, Methods, Results, and Discussions) format. This study focuses on verbs as instruments of conceptual, textual and rhetorical activation in scientific discourse and investigates the relation between the distribution of verbs, the different rhetorical moves of abstracts and the activation of conceptual areas. To that end, we propose a semantic classification of the most frequent verbs in each rhetorical section and in abstracts as a whole, and apply a method for the extraction of conceptual, textual and rhetorical information based on the analysis of a POS-tagged corpus within the framework of corpus linguistics and the Functional-Lexematic Model.

Keywords: translation, terminology, structured abstract, tagged corpus, lexical chains, semantic classification of verbs

1. Introduction

Abstracts constitute the most widely used genre of medical literature (Busch-Lauer 1995: 175). According to the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals* (International Committee of Medical Journal Editors 2006),[1] in contributions submitted to the editorial board of a journal:

> The second page should carry an abstract (of no more than 150 words for unstructured abstracts or 250 words for structured abstracts). The abstract should state the purposes of the study or investigation, basic procedures (selection of study subjects or laboratory animals; observational and analytical methods), main findings (giving specific data and their statistical significance, if possible), and

Terminology 13:1 (2007), **61–84**.
ISSN 0929–9971 / E-ISSN 1569–9994

> the principal conclusions. It should emphasize new and important aspects of the study or observations.

Despite the fact that abstracts are a much conventionalized genre (ISO 214: 1976), non-native speakers of English usually have difficulties in writing abstracts and finding the appropriate vocabulary to express these purposes, procedures, findings and conclusions. According to Busch-Lauer (1995), such difficulties are the result of a set of constraints, of which the following are relevant to terminology and translation:

1. Inadequate knowledge of the organizational principles of abstracts;
2. The fact that specific disciplines have different preferences in the use of verbs;
3. Variation in the distribution of rhetorical sections in abstracts according to the author's intention, type of research (review article abstracts, research paper abstracts or abstracts of case reports), and content of the document.

This article explores the role of verbs as instruments of conceptual, textual and rhetorical activation in scientific discourse. Using an English corpus of oncology abstracts, we examine the relation between the distribution of the most frequent verbs, and the rhetorical moves of this genre. To that end, corpus linguistics is applied to the extraction of verbs and the identification of lexical chains. We conclude that there is a correspondence between the main rhetorical moves and the conceptual areas activated in the text. Finally, a semantic classification is proposed based on the Functional-Lexematic Model (Martín Mingorance 1984, 1989, 1995; Faber and Mairal 1999).

2. Theoretical principles

2.1 Genre analysis and the rhetorical sections of research articles

In order to better understand the rhetorical sections of abstracts, i.e. the structural division of abstracts into headings (Purpose, Materials and Methods, etc.), genre analysis is particularly useful. Genre analysis has shed light on the ability of native speakers of a language to recognise a set of communicative purposes and a schematic structure prototypical of certain texts, which are associated with conventional social situations in a particular culture (Swales 1990; Bhatia 1993; Paltridge 1997). Genre analysis has put a special emphasis on the notions of *move* and *step*, coined by Swales to refer to the rhetorical segments associated with content and communicative functions, which are prototypical of certain genres and can be recognized by the members of a discourse community. Each move consists of several steps, which may be obligatory or optional. For instance, research article

introductions include three moves (Swales and Feak 1994: 175). Move 1 (Establishing a research territory) encompasses two steps: first, showing that the research

Table 1. Moves and their constituent elements (submoves) in research articles (Nwogu 1997: 125)

Move	Discourse function[2]	Section
1	Presenting Background Information	Introduction
	a. Reference to established knowledge in the field	
	b. Reference to main research problems	
2	Reviewing related Research	
	a. Reference to previous research	
	b. Reference to limitations of previous research	
3	Presenting New Research	
	a. Reference to research purpose	
	b. Reference to main research procedure	
4	Describing Data Collection Procedure	Methods
	a. Indicating source of data	
	b. Indicating data size	
	c. Indicating criteria for data collection	
5	Describing Experimental Procedure	
	a. Identification of main research apparatus	
	b. Recounting experimental process	
	c. Indicating criteria for success	
6	Describing Data-Analysis Procedure	
	a. Defining terminologies	
	b. Indicating process of data classification	
	c. Identifying analytical instrument / procedure	
	d. Indicating modification to instrument/ procedure	
7	Indicating Consistent Observations	Results
	a. Highlighting overall observation	
	b. Indicating specific observations	
	c. Accounting for observations made	
8	Indicating Non-Consistent Observations	
9	Highlighting Overall Research Outcome	Discussion
10	Explaining Specific Research Outcomes	
	a. Stating a specific outcome	
	b. Interpreting the outcome	
	c. Indicating significance of the outcome	
	d. Contrasting present and previous outcomes	
	e. Indicating limitations of outcomes	
11	Stating Research Conclusions	
	a. Indicating research implications	
	b. Promoting further research	

area is important (optional) and second, introducing and reviewing items of previous research in the area (obligatory).

Since abstracts usually follow the prototypical structure of research articles (Busch-Lauer 1995: 182), it is particularly interesting to present the schematic structure of the eleven moves that Nwogu (1997) recognises in this genre (see Table 1). Nwogu (1997: 122) defines *move* as "a text segment made up of a bundle of linguistic features (lexical meanings, propositional meanings, illocutionary forces, etc.) which give the segment a uniform orientation and signal the content of discourse in it." These moves are articulated around verbs of cognition and speech. In this paper, we use the term *rethorical moves* to refer to Nwogu's moves.

Table 1 suggests that, if we are to study the rhetorical structure of abstracts, we have to concentrate on verbs. In fact, specialized discourse is mainly structured around verbs in the sense that specialized knowledge can be linguistically codified by verbs representing EVENTS and STATES (Faber et al. 2001: 189–193). Presumably, there will be a correspondence between the main rhetorical moves and the conceptual areas activated by verbs in the text. We understand *conceptual areas* as the meaning areas activated by the words of a text that belong to the same lexical domain (López Rodríguez 2001). As opposed to lexical domains, conceptual areas are language-independent.

Considering that former studies on the rhetorical structure of biomedical research articles have primarily focused on syntax and textual organisation (Salager-Meyer 1994; Gledhill 2000; Thomas and Hawes 1994; Busch-Lauer 1995; Nwogu 1997),[3] it is necessary to carry out lexical and terminological studies that focus on verbs.

2.2 Terminology and corpus linguistics in the description of verbs

Verbs can become specialized in the same way as nouns, and be subject to description (L'Homme 1998). The description of verbs should specify their meaning, prototypical complementation patterns, and selection restrictions for the elements that can appear in association with them (Faber and Mairal 1997: 220). The methodology of corpus linguistics has been successfully applied in the terminological description of verbs and in the extraction of specialized knowledge (Faber et al. 2001; L'Homme 2001; Pearson 1998; Pérez Hernández 2002).

In this article, we apply the *Functional-Lexematic Model* (FLM) of Martín Mingorance (1984, 1989, 1995; Faber and Mairal 1999), a model that integrates Coseriu's Lexematics (Coseriu 1977) and S.C. Dik's Functional Grammar (1978, 1989). In the FLM the description and organization of the lexicon is based on the verb and on the distinction and interaction between the syntagmatic and paradigmatic

axes, in the sense that the most prototypical verbs from the semantic point of view are those with a more general combinatory potential.

The FLM has been successfully applied to the description of the semantic architecture of the verb lexicon (Faber and Mairal 1999) in a contrastive dictionary of lexical domains in which about 8,000 verbs in English and Spanish were classified onomasiologically.[4] The method of lexical organisation followed is based on Dik's method of Stepwise Lexical Decomposition, which is compatible with Mel'čuk's (1988) Decomposition Principle,[5] as well as on definitional analysis. The FLM organises verbs onomasiologically in a hierarchy of lexical domains and subdomains. A lexical domain is a hierarchy of lexemes all of which share the same nuclear meaning. The *definiens* marks the semantic territory covered by a specific dimension and becomes the factor that determines lexical membership. Consequently, words with the same central meaning component belong to the same lexical domain or dimension. In order to obtain the meaning components of the verbs in the lexicon, the FLM elaborates structured definitions of verbs consulting the most widely used dictionaries. Each definition is conceived as a syntactic frame with slots having certain selection restrictions and default values. Each verb in the lexicon activates its own scenario, which determines its semantic participants to a certain extent. To decide the kind of frame a verb has, the definitions of several monolingual dictionaries are segmented and put under the following headings: *nuclear meaning*, *subject*, *direct object*, *adverbial modification*, *usage label* and *pragmatic information.* The different meaning components are compared in order to obtain the most appropriate definition. In the example below (Faber and Mairal 1999: 83), the word *predict* has been lexically decomposed, so that its definition consists of a nuclear word or *definiens* (*to foretell*), and one or more features which

Table 2. Definitional analysis and lexical organisation (adaptation from Faber and Mairal 1999: 83)

To say that something will happen			
foretell	to say	correctly that sth will happen in the future.	
predict	to foretell	sth accurately, based on previous knowledge/experience.	⇨ **Manner: accurately** **Cause: previous knowledge**
forecast	to predict	sth (of weather and events difficult to foretell).	⇨ **Specific argument**
prophesy	to foretell	sth especially through supernatural inspiration.	⇨ **Cause: supernatural inspiration**

differentiate it from the words semantically connected with it. The *definiens* of *predict* coincides with its immediate superordinate term (*foretell*), whose total meaning is always present in the meaning of its hyponyms (*predict, prophesy*).

Working with this bottom-up (or data-driven) type of analysis, upwards from words to concepts, the FLM arrived at the following inventory of domains that underlie the semantic architecture of the lexicon:

Table 3. Lexical domains structuring the primary lexicon (Faber and Mairal 1999: 279–293)

Existence	Change	Feeling	Light
Movement	Perception	Speech	Possesion
Position	Cognition	Sound	Action
Contact			

The FLM has also provided the linguistic background for the research projects Oncoterm (Faber and Jiménez 2002) and Puertoterm (Faber et al. 2006). In Oncoterm,[6] the analysis of terminological contexts in a corpus of medical texts within the domain of Oncology allowed us to identify and introduce the events and objects underlying each conceptual category, as well as the properties connecting them.

2.3 Lexical chains and the role of verbs in specialized discourse

The study of lexical cohesion sheds light on the way verbs contribute to the organisation of specialized texts. In López Rodríguez (2001), we applied corpus linguistics to the study of lexical cohesion in oncology texts, and demonstrated that the way in which lexical cohesion is activated in biomedical texts reflects the conceptual network underlying a specialized field. In other words, conceptual activation is linked to lexical cohesion and the activation of lexical chains in texts. A lexical chain can be defined as a sequence of single-word and/or multi-word lexical units, which are formally and semantically related, spanning a topical unit of the text (Morris and Hirst 1991: 22–23). Table 4 is an example of a lexical chain identified in an oncology text for health professionals: radiation. This lexical chain is embedded within the more general lexical chain treatment.

Table 4. Lexical chain RADIATION in a text about cancer (López Rodríguez 2001: 350)

RADIOTHERAPY ~ alone 14 ~ with curative intent 6 postoperative ~ 4 concurrent ~ 2 definitive ~ 2 primary ~ 2 chest ~ 1 external-beam ~ 1 Interstitial ~ 1	65
FRACTIONATION - Fractionation 5 Conventional ~ 2 ~ schemes 2 standard ~ 1 - Fractions 2 - Hypofractionated schemes 1 - Fraction 1	9

IRRADIATION - Chest ~ 3 - Postoperative ~ 2 - Whole-brain ~ 2	9
BRACHYTHERAPY	4
EXTERNAL-BEAM (~ radiotherapy)	1
INTERSTITIAL (~ radiotherapy)	1
TOTAL	89

3. Methodology

Our methodology integrates corpus analysis and the study of lexical cohesion in order to gain a deeper knowledge of the meaning of the most frequent verbs in the different rhetorical sections of medical discourse, and to better understand the process of scientific communication.[7]

3.1 Corpus selection and description

Our corpus contains abstracts about oncology contained in the bibliographical database MEDLINE (National Library of Medicine 2006) between 1998 and 2001 to ensure quality in terms of content and relevance.[8] According to Vandaele (2001), bibliographical databases such as MEDLINE constitute an important source of information for medical translators, both at the terminological and conceptual levels. One hundred and fifty-six abstracts have been selected at random among those abstracts fulfilling the following criteria:

– Containing the word *neoplasms* as keyword or Medical Subject Headings (MeSH) (National Library of Medicine 2006);
– Summarizing the content of an article originally published in English;

- Being contained in an article published in a country where English is the official language or one of the official languages, in particular, Australia, Canada, the United States, Ireland, New Zealand, the United Kingdom and South Africa;
- Being structured into sections (PURPOSE, BACKGROUND, etc.);
- Having been written by, at least, one author with an English first name;
- Not belonging to a review article.

The last three criteria have been used to enhance the accuracy of the results provided by the search engine upon the following search criteria:

> Search: neoplasms and (LA=ENGLISH) and ((CP=AUSTRALIA) or (CP=CANADA) or (CP=ENGLAND) or (CP=IRELAND) or (CP=NEW-ZEALAND) or (CP=SCOTLAND) or (CP=SOUTH-AFRICA) or (CP=UNITED-KINGDOM) or (CP=UNITED-STATES)) and (PT=JOURNAL-ARTICLE) and (XREC=ABSTRACT)

After random selection, two different versions of the abstracts were saved: one containing information about each text (title, authorship, full reference to the article, language, country of publication, abstract, MeSH and publication type), and a short version, made up of 43,690 words.

3.2 Tagging the corpus

This section deals with the tagging process carried out in order both to recognise the rhetorical structure of abstracts and to identify verbs semi-automatically. To that end, a Part of Speech (POS) tagger and Wordsmith Tools[9] lexical analysis software have been used.

When tagging abstracts for textual analysis, we tried to maintain the variety found in the titles of the different sections, while applying unifying tags. These tags permitted the automatic recognition of four main rhetorical sections in abstracts.

The unifying tags are placed at the beginning of each abstract (<A>), of the section that deals with the procedures used (<METHODS:>), of the <RESULTS:> section, and of the <CONCLUSIONS:> section. Each tag is associated with a closing tag (</A>, </METHODS:>, etc.). We added other tags corresponding to other sections which, despite being less frequent, illustrate the stylistic variety accepted within the conventional structure of abstracts. These tags include <PURPOSE>, <BACKGROUND> and <MATERIALS AND METHODS>.

In the part of speech recognition process of our corpus, we used the tagging service of the Natural Language Processing and Information Retrieval Group of the UNED (Spanish National Distance University),[10] which is based on the well-

known Brill POS Tagger (Brill 1993). This program automatically assigns the following grammatical categories:

Table 5. Part of speech tags assigned by the Brill POS Tagger

CC	coordinating conjunction
CD	cardinal number
DT	determiner
EX	existential 'there'
FW	foreign word
IN	preposition or subordinating conjunction
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	modal
NNP	noun, singular or mass
NNPS	noun, plural
TO	infinitive marker 'to'
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, gerund or present participle
VBN	verb, past participle
VBP	verb, non-3rd person singular present
VBZ	verb, 3rd person singular present
WDT	wh-determiner
WP	wh-pronoun
WP$	possessive wh-pronoun
WRB	wh-adverb
XNOT	not and n't

As shown in Table 6, the tagged version of the text visually hinders the analysis. As a result, those tags that identify parts of speech other than the verb have been eliminated with Wordsmith Tools' Text Converter. More specifically, all tags except for those beginning with *VB* and those corresponding to modal verbs (*MD*) have been cut. After this process, the text is ready for analysis and has a simpler layout (Table 6):

Table 6. Tagged text and final layout of the texts in our corpus

(a)
<A> <PURPOSE:> The/DT long/JJ natural/JJ history/NN of/IN early/JJ stage/NN prostate/NN cancer/NN is/VBZ well/RB recognized/VBN and/CC a/DT conservative/JJ approach/NN to/TO the/DT treatment/NN of/IN elderly/JJ men/NNS is/VBZ often/RB encouraged/VB .

(b)
<A> <PURPOSE:> The long natural history of early stage prostate cancer is/ VBZ well recognized/ VBN and a conservative approach to the treatment of elderly men is/ VBZ often encouraged/ VB.

3.3 Extracting frequency lists and concordances in the different rhetorical sections

In order to obtain specific data for each rhetorical section, the settings of Wordsmith Tools were adjusted so that the program read only the parts of the file delimited by the tags associated with a rhetorical section (for instance, <RESULTS:> and </RESULTS:>). In this way, the corpus is divided into four parts: an Introduction subcorpus with 8,245 words, one for the Methods section of 10,427 words, one for the Results section (14,751) and one for Conclusion/Conclusions (5,756).

Afterwards, in the Concord function of Wordsmith Tools, we entered *VB** in the search box and re-sorted the concordances according to the second word to the left (2L) of *VB** then according to the first word on the right (1R), and finally according to file.

Concordances facilitate disambiguation in cases of homonymy or polysemy, since it is possible to arrange the lines according to the co-text of the word in question, so that its different uses are displayed. The study of the different uses of a form contributes to grasping its meaning(s). For instance, we used concordances to differentiate between the cases in which the verbs *be* and *have* function as auxiliaries. To that end, we re-sorted the concordances around instances of the verb *be* or *have* according to the second word on the right (2R) of the node *VB**. Then, we eliminated those concordances where the position 3R (in negative uses) or 2R was occupied by */VBN* (past participle) or */VBG* (present participle) as seen below:

```
1  shown /VBN that both HA and HAase are /VBP associated /VBN with bladder cancer
2  VBZ limited curative options that are /VBP associated /VBN with a high complica
3 ast cancer in the decade 1989-1998 are /VBP described /VBN. </BACKGROUND:> <METH
4  screening in TSBC's referral area are /VBP discussed /VBN. </BACKGROUND:> <METH
5 metastatic cells. In-transit nodes are /VBP found /VBN between the primary melan
6  ich frozen /VBN section specimens are /VBP harvested /VBN can/MD be /VB haphaza
7  to the management of renal tumors are /VBP being /VBG studied /VBN intensively
8 tastases is /VBZ controversial. We are /VBP reporting /VBG our preliminary resul
9 e racial incidence of this disease are /VBP not known /VBN but most surveys sugg
```

We also eliminated the past participles forming some phraseological units included in the oncology database of Oncoterm, such as the ones containing the lexemes *advanced, localized* and so on.

```
1 inical findings are /VBP suspicious for localized prostate cancer. We describe /
2  and so they may/ MD be /VB amenable to localized therapy. </PURPOSE:> MATERIALS
```

In order to identify the most representative lexemes, lemmatised frequency lists were generated and a selection was made to include only those with a frequency

equal or superior to 0.04% in the subcorpus of the rhetorical section under study. The threshold was set at this level so that the lexical chains of each rhetorical section included verbs that occurred at least twice in the corresponding subcorpus.

3.4 Semantic classification of verbs associated with rhetorical sections and generation of lexical chains

Once the corpus was tagged and lemmatised wordlists and concordances were extracted, we did a semantic classification of the verbs found in the different sections of abstracts. For this, we applied both the FLM and the lexical domains of Faber and Mairal (1999), and the methodology described in López Rodríguez (2001: 243). In particular, we exported the data of frequency lists created with Wordsmith Tools to a Microsoft Excel spreadsheet, so that semantically related lexemes in a rhetorical section are placed in the same column. By semantically related lexemes, we mean verbs whose structured definitions indicate meaning affinity with one of the 13 lexical domains of Faber and Mairal (1999). These domains constitute an inventory of the states and events of human experience. When in doubt, concordances have been used to decide the conceptual domains into which the verbs fit. The result is a set of columns; each column represents one lexical chain of meaning-related verbs, and therefore one of the conceptual areas or domains activated in each rhetorical section.

Then, in each rhetorical section, a percentage is attributed to each lexical chain, according to the ratio between the overall number of lexemes belonging to such lexical chain and the number of words included in this specific rhetorical section.

Let us illustrate this procedure with an example (Table 7). The lexemes related to the concept COGNITION in a particular rhetorical section (for example, the Introduction) are grouped in the same column, constituting the lexical chain COGNITION. The sum of the values for lexemes semantically related to COGNITION with a frequency equal or superior to 0.04% represents 1.54% of all the lexemes in the Introduction, therefore marking the relevance of the conceptual area COGNITION in the Introduction of abstracts.

The histogram with the percentages initially attributed to each conceptual area will illustrate the conceptual frames activated in each rhetorical section of the text, as will be seen in the following section. We will also see the lexemes (grouped into lexical chains) which activate those conceptual areas in abstracts as a whole.

Table 7. Activation of terms related to COGNITION in the Introduction section (Microsoft Excel)

LEXICAL CHAIN COGNITION (INTRODUCTION)					
EVALUATE	0.37	KNOWN	0.08	CORRELATED	0.05
DETERMINE	0.30	INVESTIGATED	0.08	RECOGNIZED	0.05
ASSESS	0.18	ESTABLISHED	0.05	RELATED	0.05
ASSOCIATED	0.12	STUDIED	0.06		1.54

4. Results

4.1 Conceptual activation associated with rhetorical sections

By following the aforementioned methodology, it is possible to take an initial 'snapshot' of the conceptual organization of each rhetorical section. From this starting point, we can establish a list of prototypical verbs used in the rhetorical sections of research articles.

We have observed a clear differentiation between the conceptual activation of the various rhetorical sections. These are shown in Table 8, and in histograms related to each case of rhetorical organization.

Table 8. Conceptual activation of verbs in the rhetorical sections of abstracts

Introduction (%)		Methods (%)		Results (%)		Conclusion (%)	
1. COGNITION	1.54	1. ACTION	1.63	1. EXISTENCE	2.13	1. EXISTENCE	1.9
2. PERCEPTION	0.92	2. PERCEPTION	0.8	2. PERCEPTION	0.86	2. PERCEPTION	0.88
3. EXISTENCE	0.6	3. COGNITION	0.7	3. COGNITION	0.58	3. COGNITION	0.62
4. ACTION	0.5	4. EXISTENCE	0.54	4. POSSESSION	0.54	4. SPEECH	0.62
5. SPEECH	0.36	5. CHANGE	0.49	5. ACTION	0.42	5. ACTION	0.54
6. CHANGE	0.36	6. SPEECH	0.37	6. MOVEMENT	0.14	6. POSSESSION	0.49
7. POSITION	0.2	7. MOVEMENT	0.24	7. POSITION	0.05	7. CHANGE	0.38
8. POSSESSION	0.2	8. POSSESSION	0.19	8. SPEECH	0.05	8. POSITION	0.15
		9. POSITION	0.08	9. CONTACT	0.05		

4.1.1 *Conceptual activation in the Introduction*

In the Introduction, which in abstracts is known variously as *Introduction, Purpose, Objective(s), Context-Objective* and *Background and Aims,* there is a clear tendency towards the use of verbs of COGNITION, followed by verbs of PERCEPTION, EXISTENCE, ACTION, SPEECH and CHANGE.

In practice, in the Introduction we find the presentation of a situation within which a series of hypotheses will be studied based on the perception of certain phenomena. For this purpose, authors refer to the previous findings of experts in the field, establish the aims of the Introduction and explain the procedure which

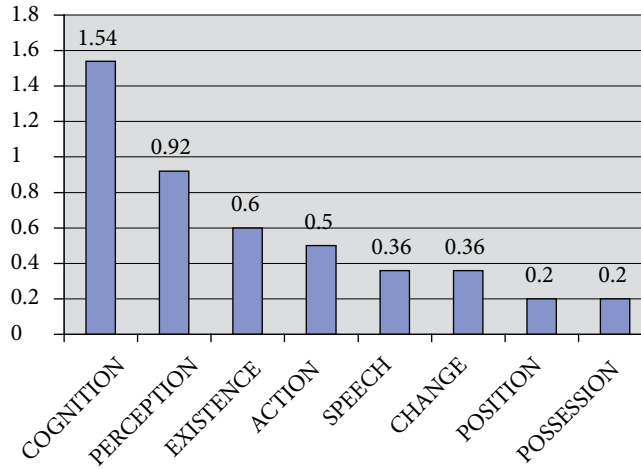

Figure 1. Conceptual activation in the Introduction of abstracts

will be followed in the study. The following verbs have a frequency in Introductions of over 0.1%: *evaluate*, *be*, *determine*, *may*, *use*, *assess*, *identify*, *detect*, *have*, *compare*, *perform*, *associate*, *increase*, *can* and *undergo*.

4.1.2 *Conceptual activation in the Methods section*

The Methods section is headed by one of the following titles: *Method*, *Methods*, *Study Design*, *Materials and Methods*, *Design/Method*, *Design-Setting-Patients-Measurements*, *Material-Study Design and Methods*, *Patients and Methods* and *Experimental Design.* The most representative lexical chain is ACTION, followed by PERCEPTION, COGNITION, EXISTENCE and CHANGE. Verbs of ACTION (*using*,

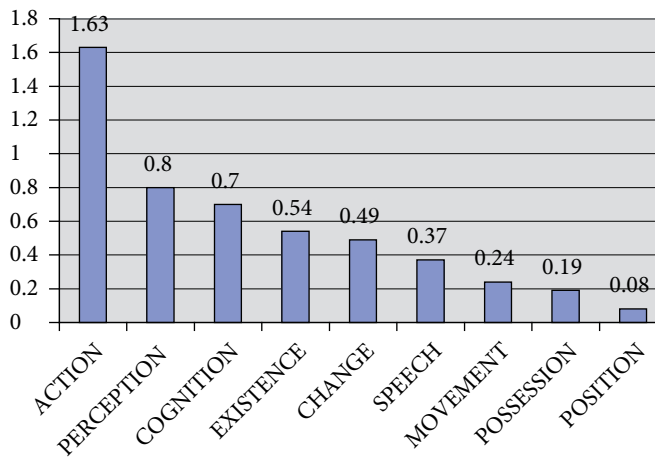

Figure 2. Conceptual activation in the Methods section of abstracts

performed, *measured*, *treated*, *completed*, *made*, *operated*, etc.) and CHANGE are especially typical and form chains which occupy a secondary position in other sections (ACTION is in 5th place and CHANGE is in 6th or 7th place in Results, and sometimes does not even appear in this section).

This indicates that the aim of the Methods section is to describe the procedures used in the gathering of data, in the actual experiment and in data analysis, and also to describe which phenomena are perceived through the experiments which have been carried out.

Among the most commonly used verbs in this section we have found in representative order: *use*, *undergo*, *perform*, *be*, *review*, *analyse*, *determine*, *have*, *compare*, *follow*, *measure*, *assess*, *receive*, *identify*, *treat*, *include* and *evaluate.*

4.1.3 *Conceptual activation in the Results section*

The Results section normally begins with the expressions *Results* or *Experience and Results.* As the rhetorical moves of this section present research observations (both those that fit in with initial hypotheses and those that do not), there is a predominance of lexical chains of EXISTENCE, PERCEPTION, COGNITION and POSSESSION. The lexical chain of POSSESSION includes verbs that link observed phenomena and characteristics attributed to these phenomena.

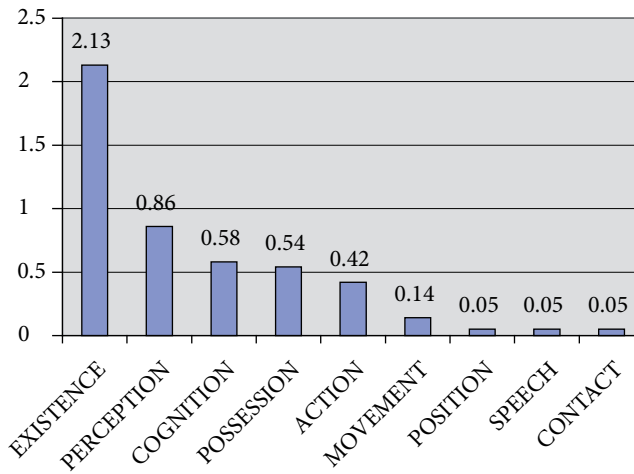

Figure 3. Conceptual activation in the Results section of abstracts

These lexical chains are mainly composed of the following verbs: *be*, *have*, *compare*, *show*, *increase*, *associate*, *find*, *observe*, *detect*, *follow* and *use.*

4.1.4 *Conceptual activation in the Conclusions*

The aim of the section called *Conclusion* or *Conclusions* is to present and to interpret the results of research carried out, and once again there is a predominance of verbs of EXISTENCE, PERCEPTION and COGNITION, which also occurs in the Results section. In addition to these conceptual areas, in the Conclusions section we also find the lexical chain SPEECH.

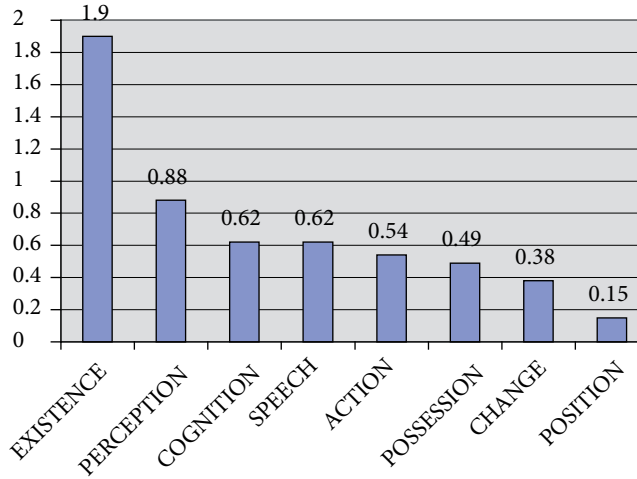

Figure 4. Conceptual activation in the Conclusions section of abstracts

The following verbs have a frequency in the Conclusions section of over 0.1%: *be*, *may*, *have*, *can*, *suggest*, *should*, *need*, *use*, *increase*, *indicate*, *provide*, *compare*, *confirm*, *detect*, *know*, *perform* and *undergo.*

4.1.5 *Evaluation of results*

We can now see if our findings coincide with one of the most widely accepted models of the conceptual macrostructure of research articles, namely the "hourglass diagram" designed by Hill et al. in 1982 (see Swales 1990: 133). According to this model, the Introduction progresses from the general context of the experiment to the particular context of the experiment in question and continues in that direction (Methods and Results), until reaching the Conclusions section. This section has a symmetrical relation to the Introduction, given the fact that there is a move from specific findings to the implications that this particular piece of research has for a more general context. In our research, the Introduction and the Conclusions section are also symmetrical: in terms of percentage, first place goes to verbs of COGNITION in the Introduction, followed by verbs of PERCEPTION and, in third place, verbs of EXISTENCE; this tendency is inverted in the Conclusions section where we can find verbs of EXISTENCE in first place, verbs of PERCEPTION in

second place and verbs of COGNITION in third place. Perhaps this is the case in the Conclusions sections because the empirical data provided by the research allows us to move from mere hypotheses (COGNITION) to observed facts (PERCEPTION).

In a similar vein, in relation to complete research articles, Swales (1990: 173) states: "Discussions, in strict contrast to Introductions, move during a cycle in an 'inside-out' direction; they move from stating the results themselves, to placing them within the established literature, to reviewing their general significance." In other words, there is a move from the inside (COGNITION) to the outside (PERCEPTION).

4.2 Lexical chains activated by verbs in abstracts

As we have already observed, in terms of writing and understanding research articles, it is of the utmost importance to recognize the conceptual areas activated by verbs in specialized discourse. Setting aside the rhetorical sections of abstracts and considering our collective corpus of abstracts related to oncology, we now present the lexical chains that are activated in order of importance. We specify which verbs are more frequent and give a further breakdown of these lexical chains. The most important lexical chains in our corpus are the following:

1. EXISTENCE: be, occur, induce;
2. PERCEPTION: detect, identify, compare, show, suggest, find, analyze, determine, observe, indicate, confirm, appear, reveal, present, examine, test, demonstrate, note
3. COGNITION: evaluate, associate, compare, determine, assess, relate, know, correlate, investigate, study, consider, diagnose, calculate, establish, recognize, imply, identify, demonstrate, estimate;
4. ACTION: use, perform, treat, measure, make, develop, test, support, complete, operate, conduct, cause, do, inject, undertake;
5. POSESSION: have, include, provide, receive, lose;
6. SPEECH: suggest, review, report, define, describe, record, account, recommend, ask;
7. CHANGE: undergo, increase, decrease, enhance, improve, change;
8. POSITION: remain, randomised, base, establish, surround, wait, assign;
9. MOVEMENT: follow, divide, remove, inject.

These lexical chains contain some of the keywords used by Nwogu (1997) to explain the moves present in research articles. The keywords have been underlined above and in Tables 1 and 9.

Moves in research articles are constructed around words from general language and, among these words, verbs play an especially significant role. By estab-

lishing conceptual areas and ordering the lexemes that activate them according to their frequency, we have tried to show the processes that are implicit in the construction of scientific knowledge. In order of importance, these processes are: the recognition of the EXISTENCE of phenomena through PERCEPTION, the COGNITIVE processes necessary to understand those phenomena and to sustain hypotheses, the ACTIONS that can modify the properties that those phenomena POSSESS, the VERBAL EXPRESSION of the scientific knowledge acquired, the CHANGES or alterations which take place in phenomena as a consequence of the actions of the scientist, and, finally, the absence of movement created by verbs of POSITION and the dynamism created by verbs of MOVEMENT.

We can now observe in more detail how these processes actually function, subdividing the lexical chains of PERCEPTION and SPEECH into smaller units (subchains). In order to identify lexical subchains and to define the meaning of verbs (Table 9), we have applied definitional analysis as understood in the Functional-Lexematic Model (see Section 2.2). The definitions provided come from the contrastive dictionary of lexical domains elaborated in the Research and Development project mentioned in note 4. Although this model is useful in the identification of lexical domains and subdomains, and in providing good terminographic definitions, we believe that the most prototypical verbs in general language are not necessarily the most frequent in our specialized corpus (see Table 9). For instance, some of the verbs presented as prototypical in the aforementioned contrastive dictionary of verbs based on the FLM *(perceive*, *recognise*, *notice*, *find*, *see* or *spot)* are not the most frequent in the frequency lists we have generated. Therefore, if we want to represent the textual and rhetorical behaviour of verbs in specialized texts, the distribution of verbs within the lexical subchains should be made in order of frequency of use.

In these lexical subchains, apart from identifying once again certain lexemes used by Nwogu (1997) to explain the moves present in research articles, we have also noticed the considerable presence of the lexical subchain EVIDENCE. This is a logical presence since, after all, scientific knowledge is constructed around what can be verified.

Moreover, some verbs can belong to more than one lexical chain, such as *note* (perception, speech), *examine* (perception, cognition) or *suggest* (speech, perception, cognition). This result does not invalidate the FLM or the methodology applied. On the contrary, it confirms new sociocognitive approaches to terminology. These approaches, following prototype theory (Rosch 1975), state that that "some units of understanding can be more easily defined in terms of a clear-cut intension and extension than others" (Temmerman 2000: 66) and that "some categories have fuzzy boundaries, i.e. there is no clear division between what belongs to the category and what does not" (Temmerman 2000: 64). In fact, the boundaries between

Table 9. Lexical subchains as a breakdown of the lexical chains of PERCEPTION and SPEECH

2. PERCEPTION: *(higher frequency)* detect, identify, compare, show, suggest, find, analyze, determine, observe, indicate, confirm, appear, reveal, present, examine, test, demonstrate, note (*lower frequency*)	
2.1. general perception (all senses)	detect: to perceive sth that others do not, with difficulty. find: to become aware of sth by chance/after looking for it. determine: find out or establish precisely. note: to become aware of something through your senses or experience and say it.
2.2. observation (to see by intentionally directing your eyes)	observe: to watch carefully (esp. in order to learn sth). examine: to look carefully and closely at sth.
2.2.1. observation + categorization	analyse: to examine in detail the constitution or structure of sth. identify: to recognise sth, assigning it to a certain category. compare: estimate the similarity or dissimilarity of sth.
2.3. evidence (to cause sb/sth to be seen)	show: to cause sth to be seen. suggest: to say something to somebody else to put an idea in their mind. indicate: to point to (fml). confirm: provide support for the truth or correctness of sth. appear: to come to be seen. reveal: to show sth that was hidden. present: to show sth. in public. test: to examine or try the qualities of sb or sth. demonstrate: to show clearly, proving by facts, actions, feelings, etc.
6. SPEECH: *(higher frequency)* suggest, review, report, define, describe, record, account, recommend, ask *(lower frequency)*.	
6.1. to suggest	suggest: to say something to somebody else to put an idea in their mind. recommend: to suggest that sth should be done/sb would do a job well.
6.2. to tell	review: to view retrospectively report: to tell sb information about sb/sth formally and publicly. describe: to tell sb what sth is like.
6.3. to explain	define: to explain the meaning of sth (esp. a word/phrase). account for: to explain the cause of sth.
6.4. to ask	ask: to say something in question form in order to get an answer.
6.5. to set down words	record: to set down in writing or some other form for later reference.
6.6. to say sth to draw attention to it.	note: to remark sth to draw people's attention to it.

semantic categories are fuzzy because the verbal lexicon constitutes a semantic macronet composed of "lexical connections" or "mappings" (Faber and Mairal 1999: 251). This means that verbs can belong to more than one lexical domain, and that different domains are projected onto others by means of metaphorical projections. Faber and Mairal (1999: 252–270) illustrate many domain connections such as the one between visual perception and cognition (*He contemplated the landscape; He contemplated the possibility of staying at home*) or visual perception and speech (*Mary observed her boyfriend dancing with another girl at the party; "He is certainly having a good time," she observed*). In any case, in the FLM, the superordinate of a lexical domain or subdomain is considered as the prototype of the category.

In the case of lexemes of "double membership," the concordances allow us to distinguish which conceptual areas are principally activated in specialized texts. Thus, the lexeme *suggest*, which in general language belongs to the lexical domains of SPEECH, PERCEPTION and COGNITION, in our corpus invariably refers to COGNITION (*to put an idea into somebody's mind*) and PERCEPTION (EVIDENCE), and is hardly used as a verb related to language, as we can see in the following concordances:

```
en necessary. /VB It is /VBZ suggested /VBN that genetic differences between PRC
     <CONCLUSIONS:> These data suggest /VBP that age at diagnosis, possibly rela
splasia and OSCC. The findings suggest /VBP that the level of NO produced /VBN b
  <CONCLUSIONS:> These results suggest /VBP that familial aggregation of stomach
   <CONCLUSIONS:> Our findings suggest /VBP that important differences in qualit
cer-specific death. /VBG This suggests /VBZ the potential usefulness of PSA as a
f stage T1 grade 3 disease. We suggest /VBP that recurrence should/MD be /VB det
tment of recurrent disease. We suggest /VBP better selection of patients for thi
```

The difficulty of establishing the boundaries between the semantic categories of COGNITION, PERCEPTION and SPEECH is also mentioned in previous studies carried out on verbs used for quoting prior research (reporting verbs). Thompson and Ye (1991) distinguish between "verbs referring to mental processes," "verbs referring to research processes" and "verbs referring to textual processes," categories which are quite similar to those proposed by Thomas and Hawes (1994): "cognition activity verbs," "experimental activity verbs" and "discourse activity verbs."

5. Conclusions

This article focuses on the verbs used in abstracts of research articles following the IMRAD (*Introduction*, *Methods*, *Results and Discussion*) format. Studying such

texts from a terminological perspective is extremely useful in order to gain a deeper understanding of scientific communication and translation.

In terms of writing and understanding research articles, it is of the utmost importance to recognize the conceptual areas activated by verbs in specialized discourse and with which rhetorical sections those verbs are usually associated. From the framework of the FLM of Martín Mingorance and the ONCOTERM and PUERTOTERM research projects, we have analysed the conceptual areas activated by verbs in a corpus of English abstracts, and how these conceptual areas are distributed throughout the four rhetorical sections. This type of analysis highlights the interrelations between the rhetorical function of the sections of the abstract, verb frequency, lexical chains, and conceptual activation. In the same way as the FLM's identification of lexical domains is an attempt to describe the semantic architecture of the lexicon in order to reveal how humans perceive, categorise and impose order on our impressions of the world, this paper describes how the processes involved in the construction and reporting of scientific knowledge are presented in abstracts in order to shed light on the way writers of research articles conceive these processes. Using corpus analysis, we have shown that, when reporting scientific discoveries, some processes are more important than others. We have also noted that the most frequent verbs in the lexical chains of our specialised corpus do not necessarily coincide with the most prototypical verbs of the lexical domains of the English verbal lexicon (general language) identified by the FLM.

This study also underlines how translation and terminology can be enriched by advances in discourse analysis, genre analysis and corpus linguistics. Although the application of corpus linguistics to the study of text type variation is not new (Biber 1988; Biber and Finegan 1991; Biber and Jones 2005), our research is original in the sense that discourse analysis is interpreted from a semantic perspective. As opposed to Biber's Multi-dimensional approach, where linguistic variation is analysed in terms of sets of co-occurring linguistic features (past tense, attributive adjectives, long words, etc.) which indicate text-type variation across five major dimensions,[11] our study concentrates on verbs and their meaning in a specific genre: the abstract. In any case, our methodology and our use of tags allow for further research on formal features of verbs. In fact, in López Rodríguez (2002), we demonstrated that certain verb forms are associated with specific rhetorical sections. We found a clear parallel between the Introduction and the Conclusions sections, where there was a predominance of past participles (VBN) and the present simple tense (VBP, VBZ). With regard to the Methods section, we observed a tendency similar to that in the Results section, as there was a predominance of the past tense (VBD), followed by the past participle (VBN).

Notes

* This research is part of the project PUERTOTERM: knowledge representation and the generation of terminological resources within the domain of Coastal Engineering, BFF2003–04720, funded by the Spanish Ministry of Education.

1. These are a set of guidelines for the format of manuscripts submitted to biomedical journals. Among other things, the requirements specify that manuscripts should include the following elements: a) title page; b) an abstract and key words; c) text; d) acknowledgments; e) references; f) tables and illustrations (figures); and g) legends for illustrations. The text should conform to the IMRAD format (Swales 1990), which stands for *Introduction, Methods, Results And Discussion*. For more information, visit http://www.icmje.org.

2. See Section 4.2 for an explanation of the underlining in this column.

3. In particular, these studies have focused on hedging in medical discourse (Salager-Meyer 1991, 1994), collocations in oncology research articles (Gledhill 1996; 2000), reporting verbs (Thomas and Hawes 1994), textual organization of English and German medical abstracts (Busch-Lauer 1995) and of complete research articles (Nwogu 1997).

4. This project was funded by the Spanish Ministry of Education: *Desarrollo de una lógica léxica para la traducción asistida por ordenador a partir de una base de datos léxica inglés-español-alemán multifuncional y reutilizable* (PB–94–0437).

5. A lexical unit should be defined in terms of lexical units that are semantically simpler than itself.

6. ONCOTERM (PB98–1342) was funded by the Spanish Ministry of Education.

7. We devised this methodology in 2002 (see López Rodríguez 2002). The usefulness of such procedure has been tested at the Department of Translation and Interpreting of the University of Granada by Reimerink (in press).

8. *Medline* is the most popular bibliographical database among health professionals. It is edited by the National Library of Medicine of the United States. For more information, see Vandaele (2001: 104–105).

9. Lexical analysis software developed by Mike Scott (http://www.lexically.net/wordsmith/index.html) and commercially available from *Oxford University Press*.

10. Natural Language Processing and Information Retrieval Group of the UNED (Spanish National Distance University, http://nlp.uned.es.

11. Biber (1988) identifies five major dimensions that define a different set of similarities and differences among spoken and written registers: informational versus involved production; narrative vs. non-narrative concerns, elaborated vs. situation-dependent reference; overt expression of persuasion; and abstract vs. non-abstract style.

References

Bhatia, V.K. 1993. *Analysing Genre. Language Use in Professional Settings*. London: Longman.

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. and E. Finegan. 1991. "On the exploitation of computerized corpora in variation studies." In Aijmer, K. and B. Alternberg (eds.). *English Corpus Linguistics.* 204–220. London: Longman.

Biber, D. and J.K. Jones. 2005. "Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles." *Corpus Linguistics and Linguistic Theory* 1(2): 151–182.

Brill, E. 1993. *A Corpus-based Approach to Language Learning.* Doctoral thesis, University of Pennsylvania.

Busch-Lauer, I.A. 1995. "Textual organization in English and German medical abstracts." In Warvik, B., S-K. Tanskanen and R. Hiltunen (eds.). *Organization in Discourse. Proceedings from the Turku Conference. Anglicana Turkuensia* 14. 175–185. Turku: University of Turku.

Coseriu, E. 1977. *Principios de semántica estructural.* Madrid: Gredos.

Dik, S.C. 1978. *Stepwise Lexical Decomposition.* Lisse: Peter de Ridder Press.

Dik, S.C. 1989. *The Theory of Functional Grammar. Part I: The Structure of the Clause*. Dordrecht: Foris Publications.

Faber, P. and C. Jiménez Hurtado (eds.). 2002. *Investigar en terminología.* Granada: Comares.

Faber, P., C.I. López Rodríguez and M.I. Tercedor Sánchez. 2001. "La utilización de técnicas de corpus en la representación del conocimiento médico." *Terminology* 7(2): 167–197.

Faber, P. and R. Mairal. 1997. "Definitional analysis in the Functional-Lexematic Lexicographic Model." *Alfinge* 9: 217–232.

Faber, P. and R. Mairal. 1999. *Constructing a Lexicon of English Verbs*. Berlin: Mouton de Gruyter.

Faber, P., S. Montero, R. Castro, J. Senso, J.A. Prieto and P. León. 2006. "Process-oriented terminology management in the domain of Coastal Engineering." *Terminology* 12(2): 189–213.

Gledhill, C. 1996. "Science as collocation. Phraseology in cancer research articles." In Botley, S. et al. (eds.). *Proceedings of Teaching and Language Corpora*, UCREL Technical Papers, Vol. 9. 108–126. Lancaster: University of Lancaster.

Gledhill, C. 2000. "The discourse function of collocation in research article introductions." *English for Specific Purposes* 19: 115–135.

Hill, S., B. Soppelsa and G. West. 1982. "Teaching ESL students to read and write experimental research papers." *TESOL Quarterly* 16(3): 333–347.

International Committee of Medical Journal Editors. 2006. "Uniform Requirements for Manuscripts Submitted to Biomedical Journals." http://www.icmje.org. Accessed 14 December 2006.

International Standards Organisation (ISO). 2004. *Documentation: Abstracts for publications and documentation* (ISO 214 : 1976). Geneva: ISO.

L'Homme, M.C. 1998. "Le statut du verbe en langue de spécialité et sa description lexicographique." *Cahiers de lexicologie* 73(2): 61–84.

L'Homme, M.C. 2001. "Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe." Presentation at the symposium *L'impact des nouvelles technologies sur la gestion terminologique*. Glendon College, Toronto, Canada, 18 August 2001.

López Rodríguez, C.I. 2001. *Tipologías textuales y cohesión en la traducción biomédica inglés-español: un estudio de corpus*. Granada: Editorial Universidad de Granada. http://www.ugr.es/~clarailr/tesis.zip. Accessed 21 December 2006.

López Rodríguez, C.I. 2002. "Extracción de información conceptual, textual y retórica en terminología: la distribución de verbos en los resúmenes de artículos experimentales." In P. Faber and C. Jiménez (eds.). *Investigar en terminología*. 167–195. Granada: Comares.

Martín Mingorance, L. 1984. "Lexical fields and stepwise lexical decomposition in a contrastive English-Spanish verb valency dictionary." In Hartmann, R.K.K. (ed.). *LEXeter 83: Proceedings of the International Conference on Lexicography*. 226–236. Tübingen: Niemeyer.

Martín Mingorance, L. 1989. "Functional Grammar and Lexematics." In Tomaszczyk, J. and B. Lewandowska (eds.). *Meaning and Lexicography.* 227–253. Amsterdam/Philadelphia: John Benjamins.

Martín Mingorance, L. 1995. "Lexical logic and structural semantics: Methodological underpinnings in the structuring of a lexical database for natural language processing." In Hoinkes, U. (ed.). *Panorama der Lexikalischen Semantik*. 461–474. Tübingen: Gunter Narr.

Mel'čuk, I. 1988. "Semantic description or lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria." *International Journal of Lexicography* 1(3): 165–188.

Morris, A.J. and G. Hirst. 1991. "Lexical cohesion computed by thesaural relations as an indicator of the structure of text." *Computational Linguistics* 17(1): 21–48.

National Library of Medicine. 2006. Medline. http://www.nlm.nih.gov/databases/databases_medline.html. Accessed 15 December 2006.

National Library of Medicine. 2006. Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh. Accessed 15 December 2006.

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

Oncoterm. 2006. www.ugr.es/local/oncoterm. Accessed 14 December 2006.

Paltridge, B. 1997. *Genre, Frames and Writing in Research Settings.* Amsterdam/Philadelphia: John Benjamins.

Pearson, J. 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Pérez Hernández, M.C. 2002. *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. Estudios de Lingüística Española, vol. 18. Madrid: CSIC/Elies. http://elies.rediris.es/elies18/. Accessed 14 December 2006.

Reimerink, A. in press. *Redactar y traducir artículos de investigación: un programa de software.* Doctoral thesis, Universidad de Granada. Granada: Editorial Universidad de Granada.

Rosch, E. 1975. "Cognitive representation of semantic categories." *Journal of Experimental Psychology* 104: 573–605.

Salager-Meyer, F. 1991. "Hedging in medical discourse: 1980–1990." *Interface* 6(1): 33–54.

Salager-Meyer, F. 1994. "Hedges and textual communicative function in medical English written discourse." *English for Specific Purposes* 13(2): 149–170.

Scott, M. 2006. WordSmith Tools. http://www.lexically.net/wordsmith/index.html. Accessed 14 December 2006.

Swales, J.M. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J.M. and C.B. Feak. 1994. *Academic Writing for Graduate Students.* Ann Arbor: University of Michigan Press.

Temmerman, R. 2000. *Towards New Ways of Terminology Description*. Amsterdam/Philadelphia: John Benjamins.

Thomas, S. and T.-P. Hawes. 1994. "Reporting verbs in medical journal articles." *English for Specific Purposes* 13(2): 129–148.

Thomson, G. and Y. Ye. 1991. "Evaluation in the reporting verbs used in academic papers." *Applied Linguistics* 12(4): 365–380.

Vandaele, S. 2001. "Utilisation des bases de donnés bibliographiques spécialisées en traduction médicale." *META* 46(1): 104–116.

Author's address

Clara Inés López Rodríguez
Departamento de Traducción e Interpretación
Universidad de Granada
Buensuceso 11
Granada 18071
Spain

clarailr@ugr.es

About the author

Clara Inés López Rodríguez teaches scientific and technical translation at the Faculty of Translation and Interpreting of the University of Granada (Spain), where she is a tenured professor and Vice-dean of Students and Work placement. She holds degrees from the University of Granada and Portsmouth. Her PhD thesis is on the relation between lexical cohesion, text type and medical translation. Her current research deals with scientific translation, the application of corpus linguistics and e-learning to translation, and cultural studies.